Big Data Approaches to Microbial Genomics

A thesis submitted for the degree

Doctor of Philosophy

Nicolas Arning

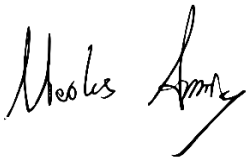
St. Hilda's college

University of Oxford

Trinity 2022

Declaration

I, Nicolas Arning, declare that this thesis was composed by myself and that the work contained herein is my own except where explicitly stated in the text. This work has not been submitted for any degree or professional qualification except as specified.



Nicolas Arning

Trinity 2022

Acknowledgments

First, I would like to thank my supervisor Daniel J. Wilson. His professional mentorship and personal support have made my DPhil as enjoyable as it was, all external factors considering. I am looking forward to hopefully promoting him from supervisor to full time friend with the hand in of this thesis.

I would also like to thank David A. Clifton as my co-supervisor, who's input was tremendously valuable, and meetings were always enjoyable. I have had the rare fortune of having two extremely friendly and personable supervisors.

Further I would like to thank my work group in Sarah Earle, Justine Rudkin, Steven Lin, Mo Yin, Jacob Armstrong, David Shaw, and Matthew Moore, for their input to my research and overall being great friends at work.

I would further like to thank my group of close friends. It is hard to find friends to rely on when it matters, but I have been fortunate to count many among them. I am indebted to Timo Kube, Jan Freese and Marius Pröve who have always stood by my side without ever having to be asked. Further my German friends and family, Saskia and Roman Assauer, Frederic Buse, Florian Röwert, Anni Dieling, Phil Knade, Patrick Kastner, Tobias Schäfers, Nicolas Mennicken, Lukas Zimmermann, Jaqui Zipfel, David Kinzler, my English friends Felix Homma, Niels Leadholm, Nasir Ahmad, Ailish Saker, Christian Peters, Emma Skeels, James Grundy, Jessica Steinberg, David Dipeolu, Jan Eijking, Geertje Bol, Kilian Kamkar, Grigalius Taujanskas and Amelia Hassoun, the DTP cohort, Oxford Basketball and all of the Selwoods. Without the support of my friends, I would not be here.

I would like to additionally thank Alexandra Casey, Nasir Ahmad, James Grundy and Grigalius Taujanskas for proof-reading parts of this thesis. I would also like to thank Samuel K. Sheppard and Aiden Doherty for helpful comments and enjoyable discussion regarding the thesis.

Finally, I would like to thank my father. His guidance, counsel, love, and sacrifice are the reasons that I can present this work. The following thesis, which I would like to dedicate to him, is as much his accomplishment as it is mine. I am eternally grateful to the biggest supporter, teacher, and friend I have ever been blessed with knowing. Thank you.

Publications and Contributions

The following paper arose from this thesis and is presented in Chapter 2:

Arning N, Sheppard SK, Bayliss S, Clifton DA, Wilson DJ (2021)

Machine learning to predict the source of Campylobacteriosis using whole genome data. PLoS Genet 17 (10): e1009436. https://doi.org/10.1371/journal.pgen.1009436

The publication is restricted to the content of Chapter 2. The work has been rewritten and restructured from the publication, with additional sections added.

The research underlying Chapter 4 was conducted as the group moved to COVID-19 research during the pandemic as part of the COVID-19 Host Genetics Initiative, which is why the title Microbial Genomics is slightly inexact with the inclusion of the COVID-19 work. In Chapter 4 the section 4.2.5 was entirely written and conceived by Daniel J. Wilson and only rewritten by me, except for subsection 4.2.5.3 which was both written and edited by me.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

Abbreviation	Explanation
AUC	**A**rea **U**nder receiver operator **C**urve
ACE2	**A**ngiotensin-**C**onverting **E**nzyme-2
BMA	**B**ayesian **M**odelling **A**veraging
BMI	**B**ody **M**ass **I**ndex
CC	**C**lonal **C**omplex
cgMLST	**c**ore **g**enome **M**ulti-**L**ocus **S**equence **T**yping
COVID-19	**C**oronavirus **D**isease 2019
FIM	**F**isher **I**nformation **M**atrix
FQ	**F**luoro**q**uinolone
GWAS	**G**enome-**W**ide **A**ssociation **S**tudy
ICD	**I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems
LMM	**L**inear **M**ixed **M**odel
LSTM	**L**ong **S**hort-**T**erm **M**emory
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MLST	**M**ulti-**L**ocus **S**equence **T**yping
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
PoP	**Po**sterior inclusion **P**robability
qq-plot	**q**uantile-**q**uantile-plot
ReLU	**Re**ctified **L**inear **U**nit
RNN	**R**ecurrent **N**eural **N**etwork
SARS-CoV-2	**S**evere **A**cute **R**espiratory **S**yndrome-**Co**rona**v**irus-2
SNP	**S**ingle **N**ucleotide **P**olymorphism
ST	**S**equence-**T**ype
UKB	**UK B**iobank
WGS	**W**hole **G**enomic **S**equences

Abstract

Alongside tremendous challenges in infectious diseases, like the rise of antimicrobial resistance and the coronavirus disease pandemic, the 21st century is also witness to the big data revolution, which offers opportunities to design methodology capable of addressing these great challenges. Whilst developing tools there are two competing philosophies of how to gain insight from big data: The modelling approach, where the natural data generating mechanism is approximated by statistical inference, and the algorithmic approach, where general-purpose algorithms are tuned to capture hidden structure in the data for prediction. The aim of the thesis is to contribute existing infectious disease problems, by motivating, designing, and applying the correct big data methodology, whilst facilitating future use through generating applications can be easily re-purposed. I first design a machine learner that can *predict* the source of Campylobacteriosis 33% more accurately than the previous most commonly used methods. Our method broadens the data input spectrum to captures of whole genomes, which uniquely allows assigning sources to individual samples showing a shift in host affinity of one of the most common lineages of *Campylobacter jejuni*. Based on the individual prediction of the machine learner,I *infer* which genetic changes are associated with host specificity by conducting a genome-wide association study. I find fluoroquinolone resistant genes pre-adapting chicken isolates to infection for humans and polyphosphate pathway associated genes to distinguish adaption to chicken and ruminant niche. For the study of COVID-19 risk, I conduct a machine learning *prediction* of very severe forms of the disease, hospitalisation, and susceptibility, whilst also

inferring risk factors for all phenotypes by applying Bayesian model averaging. I re-discover commonly defined risk factors describing socio-economic standing, ill health and ethnicity whilst discovering more novel factors like previous lung injury predisposing very severe COVID-19 and bring order to the wealth of published COVID-19 risk studies. In the closing arguments I give limitations of my work and give recommendations on how the developed tools can be re-applied to make big data research more accessible. I also expand how statistical inference and machine learning prediction can be used in unison to tap into the potential of big data to address the foremost infectious disease challenges of our time.

Chapter I

Introduction

Overview

The importance of infectious disease in addressing the major challenges of the 21st century has been made painfully obvious following the coronavirus 2019 outbreak. Alongside other lingering threats, like the rise of antimicrobial resistance, this century is also witness to a big data revolution. Alongside datasets ever-increasing in diversity, size, and speed of acquisition come unique difficulties and tremendous opportunity to circumvent and combat infection. Methods development and application have seen a parallel growth of innovation to accommodate pitfalls and harvest potential. There are two largely separate trajectories of method development which adhere to one of two competing dogmas on how to generate insight from data. Data modelling tries to approximate natural data generating processes by using knowledge to mathematically approximate the underlying sampling distribution of the data in a process known as inference. Algorithmic modelling uses data to tune a general-purpose algorithm to closely mirror the structure in the data observed in nature for the purpose of prediction. Currently inference lies within the domain of statistics with machine learning being the tool of choice for prediction, each harbouring their own strengths and limitations. In the following I will outline the differences in thinking and resulting methodology with their respective applications in infectious disease research. I aim to convey which approach is suitable for the varying challenges which infections pose in the 21st century and motivate their combination where applicable. Despite resulting from diverse philosophies of data use, and adhered to by sometimes separate communities, machine learning and statistical inference clearly have complementary and overlapping aims of knowledge discovery. The combination and eventual synthesis of statistics and machine learning holds tremendous potential at the brink of the big data era in infectious disease research.

1.1 Infectious Disease Research Enters the Big Data Era

The emergence and subsequent success of antibiotics alongside rapid advancement in public health during the 20th century led to optimism towards infectious diseases at the middle of the century (Cohen 2000). Surgeon General William H. Stewart had told the United States congress in 1969 to "close the book on infectious disease" and thus proclaimed mankind's victory over its longstanding scourge (Cohen 2000). At the turn of the 21st century the optimism had largely waned. Outbreaks of novel diseases like Ebola and Marburg haemorrhagic fevers, human monkeypox, bovine spongiform encephalopathy, severe acute respiratory syndrome, West Nile virus and avian influenza caused concern in the research community (Lashley 2006). Infectious diseases were reported to have the highest incidence and mortality of all human diseases in the last decade (Gilmour *et al.* 2013), which was prior to the emergence of coronavirus disease (COVID-19) that has dominated this decade so far. The global pandemic that followed made the importance of infection in public health abundantly evident, especially in terms of trying better to prepare for pandemics to come. Anticipating the emergence of hitherto unknown aetiological agents is difficult but epidemics that began before the emergence of severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) continue to wreak havoc. Infectious diseases like HIV and malaria are still spreading, particularly in developing countries, while past pandemics like tuberculosis are re-emerging (Laxminarayan 2022). The over- and misuse of broad-spectrum antibiotics has made antimicrobial resistance a rising concern for health care systems that just survived a flood of COVID-19 patients with a herculean effort (Ventola 2015). The concern for the resurgence of past aetiological agents and emerging future epidemics is exacerbated by

an interplay with population density (Hazarie *et al.* 2021). The largest population growth is expected to occur in areas with poor access to treatment, low urbanisation and globalisation, and subject to the worst effects of climate change and civil conflict (Bloom and Cadarette 2019). As countries in the 21st century are intertwined by international business travel, tourism, import and export, any local cause for concern can swiftly become global (Sönmez, Wiitala, and Apostolopoulos 2019). The ongoing fight against the COVID-19 pandemic, battling emerging infections in our recent past and the alarming rise of drug resistance, have painfully curbed our past hubris towards infection. While far from declaring victory over infectious diseases today, there is cause for optimism in the research community. Alongside new and demanding challenges, the 21st century has also brought new tools to combat novel and ancient pathogens.

The notion of extrapolating from the clinical past to the future has been the backbone of evidence-based medicine since its inception (Ehrenstein *et al.* 2017). The current transformation of the field comes not from a change in concept but from the variety and volume of the evidence to draw on. A long history of scientific and industrial progress culminated in the technological revolution of the 21st century (Popkova and Gulzat 2020), causing a dramatic increase in the amount, variety and speed of data suffusing all aspects of society (Dórea and Revie 2021). Albeit lagging behind progress in other disciplines like business, marketing, geography and climatology, infectious disease research has started to embrace data-driven approaches (Simonsen *et al.* 2016; Bansal *et al.* 2016). Expanding existing data streams, e.g. through better surveillance systems, and emerging novel streams, e.g. wearables, have together ushered in the "Big Data Era" for infectious disease (Simonsen *et al.* 2016). With an ever-increasing wealth

of data shifting the goalposts of what is "big", a commonly accepted definition of big data has been put forward via the three "V"s: velocity, variety and volume (Douglas 2012), with some definitions counting as many as six "V"s.

- **Velocity** describes the speed at which data is collected. In infectious diseases the acquisition of information has, for example, been accelerated by the co-option of novel data sources such as Twitter, internet search term trends and information on movement from location data.
- **Variety** of data in infectious disease research is particularly apparent in the large amount of research utilising large prospective cohort studies containing a vast array of different descriptors. Databases like UK Biobank provide phenotypic, genotypic, questionnaire and physical information alongside sample assays, accelerometery data, multimodal imaging, and comorbidity information (Sudlow *et al.* 2015).
- **Volume** of data has increased tremendously in multiple disciplines contributing to infectious diseases, spurred by technological breakthroughs. In genomics for example, the advent of high throughput technologies like next generation sequencing and microarrays has made the accumulation of large datasets economically viable (Ow, Tang, and Kuznetsov 2016). The first human draft genome was published in 2001 with an estimated cost of $300 million (Lander *et al.* 2001; Service 2006). Today the lowest price offered to sequence one human genome is $600 (Preston, VanZeeland, and Pfeiffer 2021) and the number of sequenced human genomes is expected to lie between 100 million to 2 billion by 2025 (Stephens *et al.* 2015). The enormous potential of genomics is further heightened by equally vast increases in the size of related data like gene expression, RNA and protein sequence

data, protein-protein interaction data, pathway data and gene ontology information (Kashyap *et al.* 2016).

The availability of big data in every aspect of the term offers exciting opportunities for infectious disease research but also comes with unique challenges. As an example, the increase of nucleotide sequence data summarised in Figure 1-1 makes the need for novel methodology palpable.

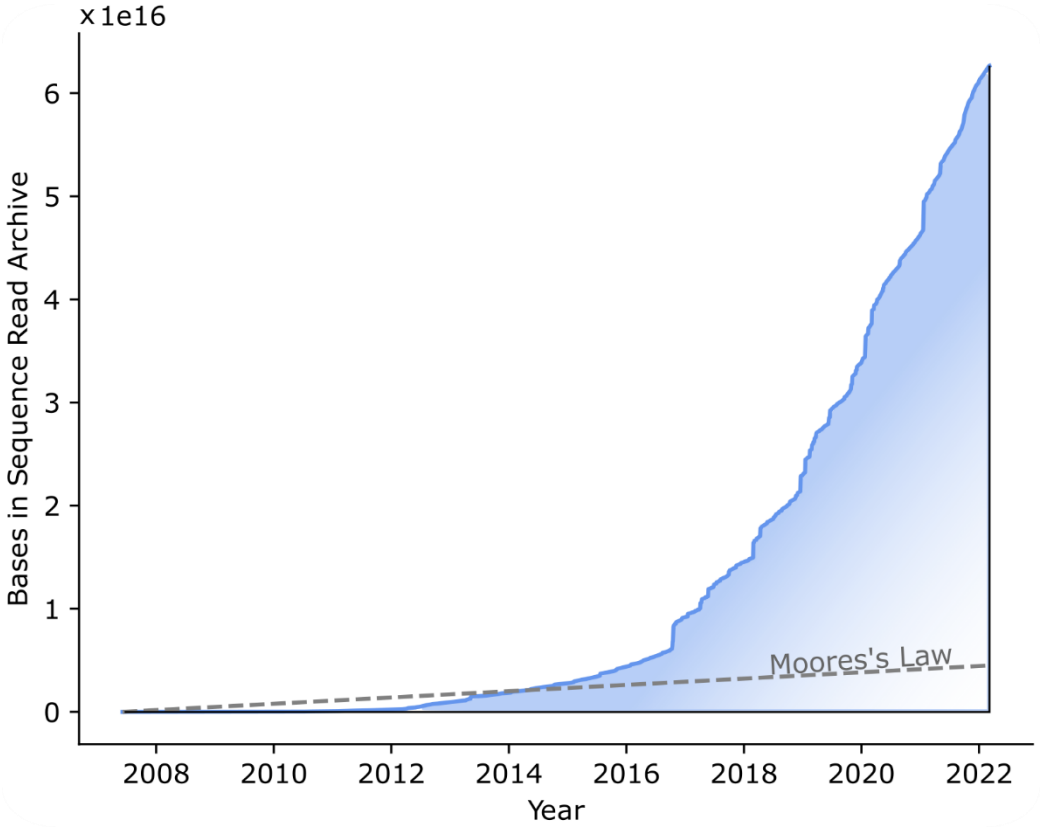


Figure 1-1 Growth of genomic data

Number of bases contained in the Sequence Read archive from the years 2008 until 2022 as registered on www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

To approximate genomic data growth, I used the number of bases contained in the Sequence Read Archive, which shows a rate of data accumulation that far outpaces

Moore's law (Moore 1965). Moore's law predicts the number of transistors per integrated circuits to double every 2 years, which has described the pace of computational innovation remarkably well for more than 50 years (Shalf 2020). If data is generated more rapidly than computational power grows then waiting for technological progress to power existing methods is a futile endeavour and demands novel approaches. The purpose of the chapter lies in providing an overview of the most common methods for using big data to generate insight into infectious disease. The subsequent individual results chapters will contain topic-specific introductions to the problems at hand.

1.2 Big Data, Big Methods

Whilst presenting novel research opportunities, big data can be challenging to incorporate into analysis owing to its complexity, size, and multiplicity of origin. The "siloing" of medical data in individual closed-off databases is still a widespread hindrance (Dórea and Revie 2021). A stipulation for successful big data research is the acquisition of actionable datasets through the combination, refinement and accumulation of data in databases (Mooney, Westreich, and El-Sayed 2015). Beyond establishing the *veracity* of the underlying data, sometimes referred to as the fourth V (Andreu-Perez *et al.* 2015), the next hurdle in extracting valuable and robust insight awaits: Scaling conventional statistical methods like regression, to high dimensional data required simplifying approximations, as many underlying assumptions are breached and heterogeneous data can be difficult to accommodate (Z. S. Y. Wong, Zhou, and Zhang 2019). Deploying big data methods, drawn from the rich toolboxes of either machine learning or statistical inference requires methods with an innate ability to use

vast and heterogeneous datasets (Ow, Tang, and Kuznetsov 2016). Machine learning algorithms mostly fall into two categories, supervised and unsupervised learning, and the combination thereof is called semi-supervised learning (Bzdok, Krzywinski, and Altman 2017). In supervised learning, labelled training data is used to learn hidden patterns to be able to accurately predict labels for previously unseen data (Greene *et al.* 2016). The focus of unsupervised machine learning also lies in pattern recognition but, in the absence of labels for the training data, is primarily used for describing data, for example identifying structure (Greene *et al.* 2016). Semi-supervised learning is based on data where some datapoints have labels that are used to predict labels for the unlabelled data (Greene *et al.* 2016). Reinforcement learning is an emerging branch of machine learning that trains an algorithm by trying to maximise the "reward" received for the performance of a given task (Sutton and Barto 2018). There are many methods within the broad classes of supervised, unsupervised, semi-supervised learning, and reinforcement learning. In this thesis, our focus lies in supervised machine learning, due to its relevance in prediction, which is of special interest in infectious disease research since it can forecast risks, phenotypes and sources of infection (Wiemken and Kelley 2020). Despite their often-superior predictive performance for non-tabular data, machine learning approaches have often been criticised for their "black box" nature owing to the difficulty of extracting actionable insight in terms of the underlying scientific processes (Wiemken and Kelley 2020). This is troublesome for medical research, especially when designing disease interventions. With both statistical and machine learning approaches enjoying respective advantages and caveats, there is no

"silver bullet" of big data analysis. Even so, infectious disease researchers have used big data successfully by leveraging the strengths of both techniques.

Considering the complete spectrum of big data infectious disease research, Iwashyna and Liu (2014) argue that there are three questions about disease that big data can answer:

- What is going to happen?
- What will happen if some variable is changed?
- What is the pattern of the underlying data?

In the following section I would like to outline milestone studies in infectious disease that address these questions using the three characteristics of big data. To answer what is going to happen in influenza outbreaks, Google Flu trends leveraged increased data *velocity*, using keyword searches and comments on social media platforms to monitor the spread of flu (H. T. Wong *et al.* 2015). The fundamental question of how changing genetic variables affects the phenotype of the organism is addressed by genome-wide association studies (GWAS) using large *volume* of genomes readily characterised by rapid high-throughput genotyping. A milestone study was the uncovering of the genetic basis of severe COVID-19 by the COVID-19 Host Genetics Initiative (2021), using genomes from 49,562 patients from 19 countries to infer 13 loci predisposing patients for severe COVID-19 upon infection. An example of discovering patterns of the underlying data is provided by Alaa and colleagues (2019) leveraging the *variety* of data contained in 473 different measurements from the 423,604 participants in the UK Biobank. The study predicts cardiovascular disease risk using an algorithm that

automatically adapts an assembly of machine learning techniques. For a more complete review of big data studies in health please see de la Torre Diez *et al.* (2016) or Doherty *et al.* (2021). The tools to answer these questions can generally be sorted into one of two categories: inference (GWAS study, Google Flu trends) or prediction (cardiovascular disease risk prediction) tracing a schism in how to use data to generate insight.

1.3 Inference Versus Prediction

Broadly speaking, big data analysis approaches can be categorised into "inference" and "prediction". Both approaches take a set of independent variables and model their relationship with the dependent variable. However, inference is concerned primarily with understanding the scientific processes by which the outcome data were generated, whereas prediction is primarily concerned with accurately imputing unseen outcomes in other datasets that are in some sense equivalent (Bzdok, Altman, and Krzywinski 2018; Wiemken and Kelley 2020). These separate but related aims of big data analysis are well-known and may be recognised under other names. For example, Breiman (2001a) defined:

- **Data modelling**, which assumes that the outcome is the result of a stochastic process that can be approximated by conceiving a mathematical model. The goal is to further the understanding of how the outcome is generated as a function of the explanatory variables. Statisticians adhere to this notion by carefully crafting bespoke probability models approximating observed outcome distributions.
- **Algorithmic modelling**, which treats the data generating process as unknown and focuses on finding a function whose output is as close to the observed outcomes as possible. Machine learning practitioners adhere to this concept by tuning a general-

purpose function with the goal of the producing an output that balances faithful reproduction of the observed outcomes in the training data, with generalisability to other, unobserved, datasets.

The concept underlying these different approaches is illustrated in Figure 1-2.

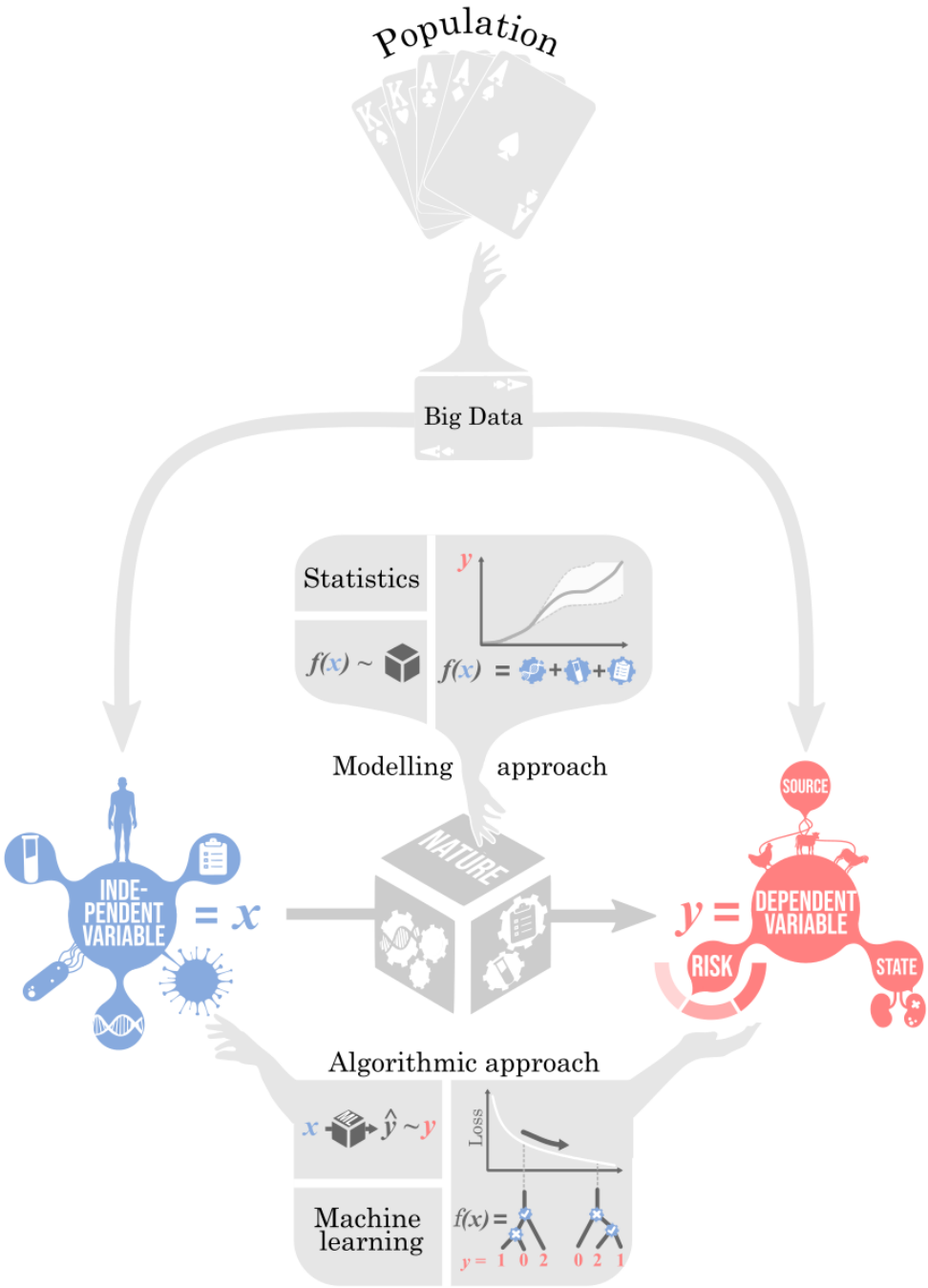


Figure 1-2 Modelling approach vs algorithmic approach

An illustration of the difference between the algorithmic approach and the modelling approach. (Big) data is a (large) sample drawn from an unknowable population. The algorithmic approach uses that sample to liken a predicted dependent variable

These concepts help us think about the different ways that researchers employ big data to generate insight: Data modelling with the aim of inference, and statistics being the tool of choice, or algorithmic modelling where the goal is to achieve the most accurate prediction through machine learning tools. Of course, there are statistical approaches to prediction, and attempts at inference through interpretable machine learning, with the distinction becoming increasingly blurred (Libbrecht and Noble 2015). The same researchers may adopt different approaches depending on their aims. Indeed, some researchers do not accept a formal distinction at all. However, comparing and contrasting both approaches shows why the same methods that accurately infer underlying processes are not necessarily equally well poised to precisely predict outcomes (Lo *et al.* 2015). Similarly, algorithms optimised for prediction may do so at the cost of interpretability (Libbrecht and Noble 2015). As the "No free lunch theorem" proclaims (Wolpert 1996) one singular model cannot be best across all problems.

1.4 Disambiguation

Before I can embark on a description of the mechanisms of inference and prediction, the terms must be coherently defined, as machine learning and statistics use different terms for the same concepts, which are summarised in Table 1-1. The *population* describes the entirety of objects from which the data at hand is a random sample.

Disambiguation

Gaining insight about the population is the aim of any data analysis, but can never be fully achieved, as sampling the population completely is infeasible. The data contains *independent variables* and *dependent variables*. Independent variables are the predictors or regressors, also known as features in machine learning. They can be split further into factors (discrete variables) and covariates (continuous variables). The dependent variables, also known as labels, targets or outcomes in machine learning, are the results of changes in the independent variables leading to changes in dependent variables. Understanding versus leveraging the relationship between the independent variables and dependent variables is fundamental for inference versus prediction respectively. As outlined in Table 1-1 I have adopted the language of statistical inference to assure consistency for this thesis.

Table 1-1 Terminology machine learning vs. statistics

Comparison of terminology and concepts that differ between machine learning and statistical inference. In this thesis I adopt the terms used in statistical inference. If there are multiple terms listed, the term I will use is highlighted in bold.

Machine learning	Statistical inference
Terminology	
Features, input	**Independent variables**, predictors, regressors, explanatory variable
Prediction, output, labels, target, outcome, classes (if factors)	**Dependent variables**, response variable
Learning	Parameter estimation
Training	Fitting a model
Generalisation	Out-of-sample prediction
Integer (if whole numbers), float	Continuous variables
Categorical variables	Discrete variables
Classes (if target of prediction)	Factors
Class (if target of prediction)	Level
One-hot-encoded variables	Dummy variables
Bias	Intercept
Recall	Sensitivity
Precision	Positive predictive value
Concepts	
Regression refers to prediction of continuous dependent variables, whereas classification predicts factors	Regression refers to prediction of both continuous and factors and dependent variables
Parameter refers to the variables that are tuned in training to learn patters of the data. Hyperparameters are variables that control the training.	Parameters are variables that describe the shape of a probability distribution

1.5 The Algorithmic Modelling Approach: Prediction Through Machine Learning

1.5.1 Algorithm Choice

Algorithmic modelling as implemented in supervised machine learning aims to make minimal assumptions, e.g. linearity, lack of multicollinearity and independence, about the data generating process (Bzdok, Altman, and Krzywinski 2018). Instead, machine learning trains general-purpose algorithms that can become arbitrarily complex to approximate a mapping function producing similar results to the structure observed in the data. The starting point of any machine learning study is to choose from pre-defined algorithms based on their empirically verified success (Bzdok, Altman, and Krzywinski 2018). There is an array of commonly used supervised machine learning techniques that have proven apt at answering infectious disease questions. I list several algorithms used for classification, but with adjustment all described methods can also be used for regression (T. Hastie, Tibshirani, and Friedman 2008).

- **K-nearest neighbours** is one of the simplest algorithms in the machine learning repertoire that classifies new data points by assigning the majority dependent variable of the k most similar datapoints in the training data (Fix and Hodges 1951). The algorithm can be provided with different distance metrics to find the k-closest neighbours (Abu Alfeilat *et al.* 2019). This is especially useful when faced with heterogeneous data which requires some custom formalisation of distance. Despite the simplicity of the concept, the k-nearest neighbour algorithm can be effectively deployed given the levels of the dependent variable to predict (or classes in machine learning) are well separated in feature space. This is demonstrated by nearest neighbour classification being successfully used to determine whether an infection is bacterial or viral with 92% accuracy from leukocyte data collected from 2,098

blood samples (Suyanto *et al.* 2020). Differentiating between infections, or sources of infection, is an important application of machine learning in infectious disease as it powers the design of targeted interventions.

- **Naïve Bayesian classifiers** are a family of classifiers that probabilistically assign values to dependent variables using an adaptation of Bayes' theorem (Xu 2018). Varying distributions, like Bernoulli, multinomial or Gaussian can be used for different features. The evidence from multiple features is combined under the assumption that all features contribute independent information about the dependent variable (John and Langley 2013). Even when the independence assumption is violated, naïve Bayes has the ability to classify competently (Webb 2010). Despite their algorithmic simplicity and strong independence assumptions, Naïve Bayesian classifiers are widely used. F. Li and colleagues (2020) use naïve Bayes to distinguish between commonly found infections from a set of 146 predictors, including incidence rates, symptoms, test results and epidemiological data. The resulting median accuracy of 97% over all 25 predicted diseases shows the competitive predictive performance achievable with the naïve Bayesian classifier.
- **Support Vector Machines** optimise a hyperplane through feature space that separates different levels of the dependent variable with maximal margins (Cortes and Vapnik 1995). In the case of linearly inseparable levels of the dependent variable, the data is transformed by mapping into higher dimensional space, where the levels are potentially distinct (McIntyre *et al.* 2017). The transformation functions are called "kernels", of which there are numerous to choose from, like polynomial or Gaussian depending on empiric classification performance (McIntyre

et al. 2017). The "kernel trick" allows for optimisation of the algorithm without operating in higher dimensional feature space which would be computationally costly (Ben-Hur *et al.* 2008). Pairwise distances between datapoints computed by the kernels can be used instead (Ben-Hur *et al.* 2008). Support vector machines have been successfully implemented to predict COVID-19 status based solely on symptom data with 87% classification accuracy (Guhathakurata *et al.* 2021). Predicting the presence or absence of disease through algorithms can support the manual decision-making process by healthcare professionals.

- **Decision Trees** partition data successively, applying a series of binary splits using individual features to produce a tree-like structure (Quinlan 1986). The "leaves" of the tree are data bins which are assigned levels of the dependent variables as labels (Quinlan 1986). The splits are selected in training to optimise a measure of choice, for example to minimise the entropy in both resulting splits (M. Li, Xu, and Deng 2019). As variables are considered individually, there is no need for standardisation of the input data (Quinlan 1986). Decision trees have been applied in the ongoing COVID-19 pandemic to predict mortality in severe cases (Yang *et al.* 2021). Assigning the risk of disease mortality or severity through machine learning is important for resource and attention allocating in the health sector. Another benefit for clinicians is that trained decision tree algorithms are easily interpretable, with a caveat being their tendency to overfit if uncontrolled (Bramer 2013). Beyond regularisation to avoid overfitting, the averaging over many decision trees in ensemble methods is often used (Schrider and Kern 2018).

- **Random Forests** are ensembles of decision trees where each individual tree is based on a bootstrap subset of the training data and a random sampling of the independent variables (Breiman 2001b). Levels of the dependent variables are assigned by "voting" of all decision trees in the ensemble. Whilst retaining the ability to handle non-standardised input data, random forests tend to be less prone to overfitting due to the embedded stochasticity in sampling (Statnikov *et al.* 2013). A random forest classifier is at the heart of the PaPrBaG classifier that can predict novel bacterial pathogens from next generation sequencing data, without relying on sequence similarity, with 88% to 93% accuracy (Deneke, Rentzsch, and Renard 2017a).
- **Gradient Boosted Trees** take a similar approach to random forests by averaging over multiple decision trees whilst introducing stochasticity, but a single decision tree is incrementally improved instead of building an ensemble in parallel (Friedman 2001). In a process called boosting, the ensembles are built by sampling at a fixed interval from the succession of iteratively improved trees (Friedman 2001). In a process termed gradient descent, the trees are improved by following a negative gradient of a pre-defined loss function by tuning parameters of the model accordingly (Mason *et al.* 1999). The gradient boosted trees algorithm, in its many highly optimised implementations, is very popular today. For example, LightGBM (Ke *et al.* 2017) was used to predict antimicrobial resistance profiles directly from the mass spectra of clinical isolates (Weis *et al.* 2022). Machine learning can be an important tool in the fight against antimicrobial resistance, due to being able to predict resistance from genetic data among other functions (Anahtar, Yang, and Kanjilal 2021).

- **Artificial Neural Networks** are inspired by the human brain and composed of multiple layers of computational units emulating neurons (Hopfield 1982). Artificial neural networks come in a variety of vastly different architectures, with all architectures containing one or more input layers that read in the independent variables and one or more output layers that generate the dependent variable with "hidden" layers in between input and output (Hopfield 1982). Every individual neuron contains a mathematical transformation of its input variables, which is commonly non-linear (Sheehan and Song 2016). There are no connections between nodes within a layer, but nodes between layers are connected so that output from one layer is used as input to the next (Sheehan and Song 2016). Independent variables thus undergo multiple successive transformations until reaching the output layer which results in the largest information processing capacity of all machine learners (Sheehan and Song 2016). There are a variety of transformation functions, network architectures, regularisation techniques, output and input layers, loss functions and optimisation techniques to choose from. In their wide variety, artificial neural networks have shown the ability to approximate myriads of different data structures (Y. Li *et al.* 2019), especially in deep learning where many hidden layers are used (Sejnowski 2018). However, due to the immense complexities that neural networks reach in order to approximate increasingly complex natural phenomena, they have often drawn criticism for their lack of interpretability. Despite obscuring the decision making process, deep learning can be very useful for infectious disease research, especially due its superior ability to accurately classify image data compared to other machine learning algorithms (Affonso *et al.* 2017).

Zhan *et al.* (2021) demonstrates this by using 3,463 computer tomography images to distinguish between four different pulmonary infectious diseases with a median area under the receiver-operator curve of 0.99.

The list of machine learning algorithms presented here is not complete as the repertoire of algorithms is constantly growing due to the speed of technological progress within the field.

1.5.2 Algorithm Training

To fit and evaluate these general-purpose algorithms to the information at hand, the data is usually split across a training set and test set. The training set is used to adjust the parameters of the chosen algorithm. The test set is used to measure predictive performance on novel unseen data, also known as the generalisation ability of the algorithm. Minimising the generalisation error is the purpose of training machine learning algorithms. Generalisation is also known as out-of-sample prediction as opposed to within-sample prediction which is measured on the training set. Most machine learning algorithms have hyper-parameters that control the learning process, which require adjusting before training. To avoid stratifying the training data further, k-fold cross validation (Stone 1974) can be used for hyper-parameter tuning. Herein, the training data is split into k groups, with k usually being set to 5 or 10. Each of the k groups is used for testing the performance given the hyper-parameter value of interest used in a model trained on the remaining k-1 groups. The average of the k performance scorers summarises the suitability of the hyperparameter value for the prediction task at hand and the process is repeated for multiple possible values. As statistical inference estimates parameters directly from the data, all data can be directly used for fitting

without the need for partitioning further. Having found suitable values for the hyper-parameters that control the algorithms learning ability, machine learning practitioners can commence the training of the algorithm.

During training, independent variables and their respective dependent variables are used to adjust parameters of the general-purpose function. By providing labelled data as input, the algorithm is trained to predict a dependent variable from the independent variables, so that the observed relationship between known dependent variable and independent variables is preserved. The difference between predicted and observed output is quantified by a loss function, that is chosen according to the type of the dependent variable. There are many types of loss function. *Cross-entropy*, for example, may be used for factors with multiple levels, and *log loss* for binary dependent variables (Q. Wang et al. 2022). By minimising loss using data to tune parameters of the model it "learns" the underlying hidden patterns of the data (Bzdok, Krzywinski, and Altman 2017). If the dependent variable is continuous, the problem at hand is considered a regression task, whereas functions with discrete output variables are considered classification tasks. In classification settings, the different levels of the dependent variables are considered classes. The knowledge about patterns within the data contained in the trained algorithms is then used to generalise beyond the training data to predict dependent variables from previously unseen independent variables (Bzdok, Krzywinski, and Altman 2017). To estimate the generalisation ability of the algorithm beyond data seen in training, the test set is used.

1.5.3 Algorithm Testing

In testing a classification algorithm, multiple different performance metrics can be chosen based on the concepts of true positives and false positives which statistical hypothesis testing also uses. Depending on which type of error is considered most costly for the prediction at hand, a range of different scorers are available which are depicted in Figure 1-3 alongside the general machine learning workflow. Testing a regression machine learner is based on quantifying the difference between the predicted and observed dependent variables. For example, the mean of the average squared distances between predicted and observed dependent variables can be used, which is known as the mean squared error. After assuring adequate performance of the machine learner on the test set, the algorithm is finalised for the application on real world problems by re-training using both test and training set.

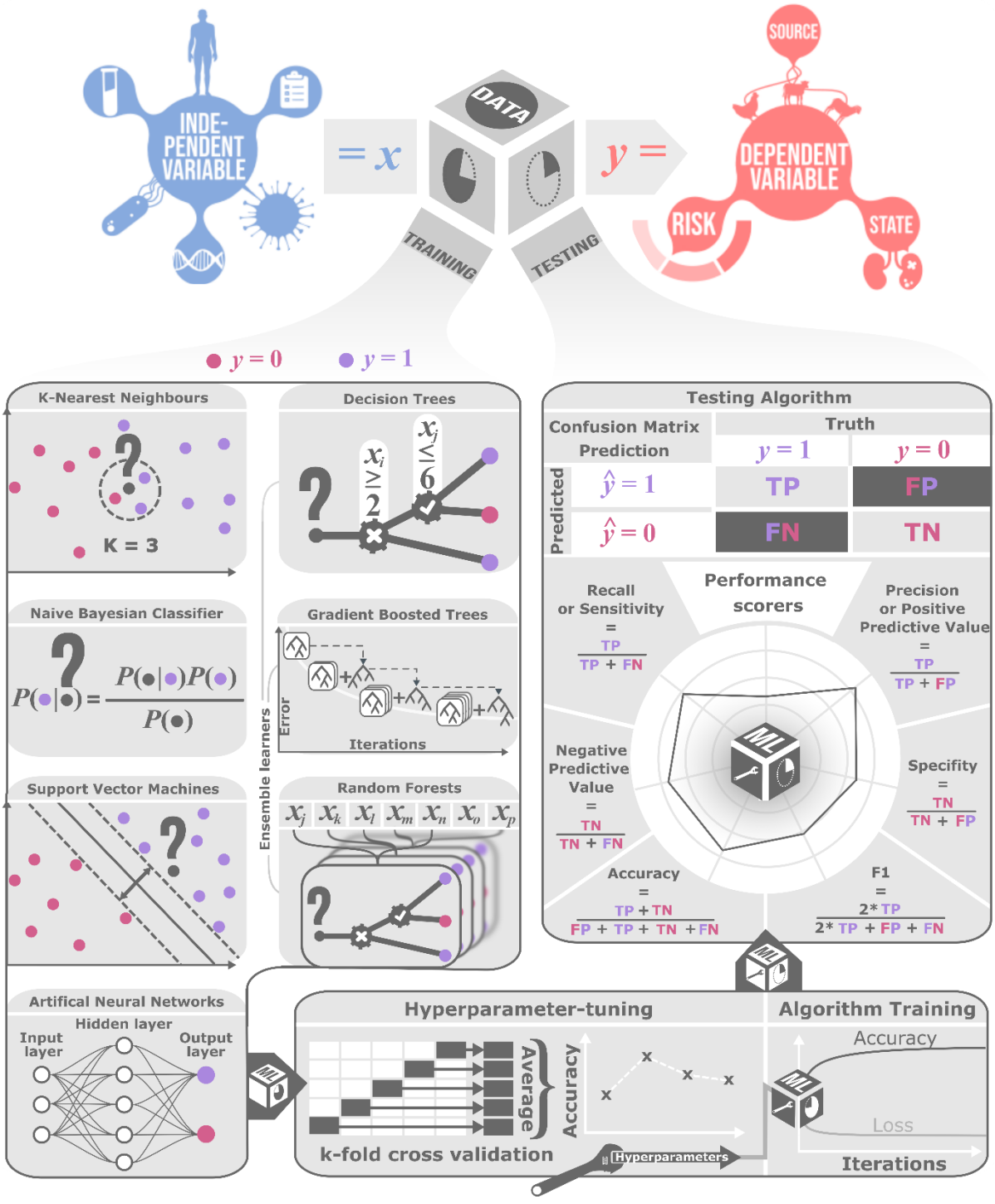


Figure 1-3 Machine learning workflow as exemplified in classification tasks

The data is split into training and testing, after which a suitable general-purpose algorithm is chosen, its' hyper parameter tuned and fitted to the training data. The performance of the fitted classifier is subsequently measured using a metric of choice.

Tempted by the demonstrated success for understanding infectious disease, machine learning practitioners must be mindful of multiple pitfalls arising due to insufficient or erroneous training. A first hurdle to overcome is underfitting where the model does not grow to sufficient complexity to adequately capture the relationship between the independent and dependent variables. Different model choice or more training data can effectively combat underfitting and rescue generalisation performance. The opposite problem is overfitting as a widespread result of the "curse of dimensionality" (Bellman 1966). Given enough variables, a model can grow so complex that alongside genuine data patterns it can also learn the stochastic noise inherent in any sample (Bzdok, Krzywinski, and Altman 2017). Whereas performance on training data is high, it is diminished in the presence of unseen data that is subject to different random measurement fluctuations, meaning the generalisation ability of the algorithm is low. Limiting the complexity of the algorithm through regularisation, or providing more training data can ameliorate the effects of overfitting (Bzdok, Krzywinski, and Altman 2017). Complex algorithms with a high capacity, like artificial neural networks, are more prone to overfitting, whereas simple learners like naïve Bayesian classifiers are more likely to underfit (Bashir *et al.* 2020). Lack of sufficient training data can also be limiting when the model encounters patterns not seen in training. Google Flu trends for example missed the first 2009 wave of the A/H1N1 influenza entirely and overstated the risk from A/H3N2 in the 2012/2013 flu season (Olson *et al.* 2013). Changes in internet search behaviour due to the reporting of the epidemic and difference in seasonality, geographical heterogeneity and age distribution have all been proposed as causes of

failure (Olson *et al.* 2013). Underfitting, overfitting and insufficient training can all be fixed within the framework of algorithmic modelling by adjusting the model hyper-parameters in cross-validation to control learning or by providing more training data. However, there are further problems that are the consequence of the central dogma of algorithmic modelling and difficult to circumvent whilst adhering to its doctrine.

1.5.4 Machine Learning's Blind Spot

In algorithmic modelling the data generating process which creates dependent variables from independent variables is treated as a black box (Breiman 2001a). Machine learning attempts to mirror nature by recapitulating the empirical structure within the data, including the patterns of co-occurrence between variables. There are many models that can achieve this and in any given scenario, some models are likely to achieve better predictive performance than others. With the ever-increasing wealth of data, an increasing number of empirical patterns can be accurately modelled *in silico*. With few assumptions about the data being made, the absence of knowledge about the underlying mechanisms is not necessarily prohibitive. Being liberated from formulating an underlying scientific process-driven model in machine learning can be considered a strength, since with minimal adjustments the same methodology can be applied to a wide array of problems (Dórea and Revie 2021).

In contrast to pre-formulated hypothesis-driven testing of association used in statistical inference, an idea explored later, machine learning can be treated as a "hypothesis generating machine" by allowing the data to reveal structures hidden within (Khoury and Ioannidis 2014). The ability to uncover relationships between variables connects machine learning to inference and attempts to draw inference from machine learning

are commonplace. Machine learning practitioners often rely on relative feature importance, which measures the relative contributions of each variable to the prediction process (Holz and Loew 1994). Relative feature importance is a by-product of training and can be implemented in many machine learning algorithms (Holz and Loew 1994). However, since the focus of algorithmic modelling is likening a predicted output to an observed output, the nature of the underlying associations is not the quantity optimised by the algorithm (Bzdok, Krzywinski, and Altman 2017). As long as the relationship can be leveraged to predict dependent from independent variables, differentiating between predictions driven by correlation versus causation is not of foremost concern in the goals of algorithmic modelling as long as the generalisation ability is not diminished (Hoffman and Podgurski 2013). The emerging field of interpretable machine learning tries to address the "black box" criticism by pushing insight into decision making beyond relative feature importance (Molnar 2020). Within interpretable machine learning there is a differentiation between intrinsic interpretability, where algorithms are constructed with transparency in mind, and post-hoc interpretability, where subsequent analysis generates insight into a predictor (Molnar 2020). Both are concerned with describing the relationships between variables, but the distinction between causal and non-causal associations is currently outside the grasp of the field (Murdoch *et al.* 2019).

Knowing whether a change in one variable causes the change in another is often important for infectious disease research when trying to combat infections by designing interventions (Wiemken and Kelley 2020). As diseases are the result of complex causal chains (Kraemer *et al.* 2001), understanding the relative effect of independent variables

is encumbered by confounders with the dependent variable (Skelly, Dettori, and Brodt 2012). Whereas known confounders are often explicitly thought of in statistical frameworks, it is difficult to control for confounders in machine learning settings (Ferrari, Retico, and Bacciu 2020). Although it is true that machine learning can readily generate hypotheses, it does not typically focus on testing those hypotheses. Some practitioners of machine learning view the field as encompassing classical and Bayesian statistical approaches, and such approaches can certainly be used for prediction. However, I distinguish them as being fundamentally focused on the process of inference, including parameter estimation in the presence of confounding and formal hypothesis testing.

1.6 The Data Modelling Approach: Statistical Inference

1.6.1 Overview statistical inference

In Breiman's dichotomy, data modelling is the driving concept in statistical inference, where the sample is used to elucidate the data generating process and thereby gain domain-specific insights through the application of probability theory (Casella and Berger 2021). Here, the dependent variables are treated as realised values of a random variable for which a model is constructed. Models can generally be divided into parametric and non-parametric. Parametric models make explicit assumptions about an underlying probability distribution, like assuming Gaussian or Poisson distributions, as that allows for convenient summary in a fixed set of parameters (Casella and Berger 2021). Non-parametric and semi-parametric models are based on the notion that the data generating process cannot be readily defined in full. They are often more complicated and may suffer a deficit in statistical power as a trade-off for making fewer assumptions (Casella and Berger 2021). There is a plethora of both parametric and non-

parametric statistical models to choose from based upon the data at hand and information about the mechanism to describe. However, in the context of data modelling, parametric models play a leading role.

An important factor is the form of the dependent variables. For instance, logistic regressions are used for binary outcomes and ordinary regressions are typically applied to continuous outputs (Pohlman and Leitner 2003). In a process akin to minimising a loss in machine learning, the data are used to estimate the parameters of the model in order to optimise a quantifiable goal, usually the likelihood of the data given the probability model (Casella and Berger 2021). The aim of statistical inference is to gain insight into the data generating process that produces the observed data. Investigating the relationships between variables lies at the heart of statistical inference, however in many situations statistics can never outright prove causality, particularly in analysing observational, as opposed to experimental, data (Pearl 2009). Instead, inference concentrates on:

- **Parameter estimation** which can be expressed as a point estimate, or an interval estimate for a parameter of the model. Parameters are components of the data generating process, which are estimated from the data. For example, estimating the coefficients of a regression can provide information of the magnitude and direction of the influence of an independent variable on the dependent variable.
- **Rejection of a hypothesis** through hypothesis testing. The fit of a mathematically formalised model of the data generating process, often framed as a pair of null and alternative hypotheses that differ in terms of their parameters, can thus be evaluated against the data. For example, rejecting the null hypothesis that a

regression coefficient is zero indicates an association between the independent variable and the dependent variable.

In practical terms, establishing the existence, magnitude and direction of an association between individual independent variables and dependent variables allows for the design of interventions aimed at alleviating the burden of infections (Bzdok, Engemann, and Thirion 2020). The statistical toolkit offers multiple different methods to perform parameter estimation or hypothesis testing, depending on the focus of the analysis; often both are pursued simultaneously. My aim in the following is not to give a comprehensive summary of all possible approaches to statistical inference, but rather to give an overview of commonly used methods in infectious disease.

1.6.2 Parameter estimation

1.6.2.1 Point estimates

Irrespective of the underlying data generating process in nature, all but the simplest of models require the estimation of parameters to adjust the model to fit the observed data (Beck and Arnold 1977). Depending upon the choice of model there are a varying number of parameters to be estimated; simple linear regressions for example are fully described by estimating the coefficient, intercept, and variance. The regression coefficient of the independent variable is an estimate of the size and direction of its effect on the dependent variable, and thus quantifies their relationship. The ambition of parameter estimation is to arrive at methods that have maximum consistency and minimal bias (Levy 2012). Consistency implies that the estimate is guaranteed to converge to the true underlying parameter as the amount of data increases, assuming (usually wrongly) that the model is strictly true (Levy 2012). The bias describes the

difference between expected value of the estimator across indefinitely many independent datasets and the "true" parameter, if such a quantity were known (Levy 2012). There are several different statistical paradigms and associated estimators under which parameters can be estimated. To name a few:

- **Method of moments** uses moments – low-dimensional summaries of the data – to calculate point estimates of the parameters (Hall 2004). Moments are mathematical properties of a probability distribution, e.g. the first moment is the expected value (mean) and the second moment is the variance. In the method of moments, the model is used to express population moments as functions of the parameters to be estimated. These are set equal to the observed sample moments, and the equations solved for the parameters of interest (Hall 2004). One use case for estimating parameters, is the ordinary least squares method to estimate the coefficients and intercept of a linear regression model. By using the least squares method, the model parameters are chosen to minimise the sum of the squared residuals. Residuals are the differences between the observed dependent variable and the dependent variable predicted by the regression from the independent variables (Boos and Stefanski 2013). Using the method of moments, the parameters can be solved for analytically and calculated using sample data in order to arrive at a point estimate for the regression coefficients and intercept parameters.
- **Maximum Likelihood** provides point estimates for parameters of a probabilistic model by maximising a likelihood function, so that the estimated parameters are those that make the observed data most probable (Cam 1990). The likelihood is defined as proportional to the probability of the data conditional on the model and

parameters (Cam 1990). In practice it is equivalent, and often more convenient, to optimise the parameters with respect to the logarithm of the likelihood (Cam 1990). In an ordinary linear regression, least squares can also be framed as a maximum likelihood routine. Logistic regression is fitted by maximum likelihood (Wright 1995), through a numerical optimisation technique (Zou *et al.* 2019). The regression coefficients of logistic regression can be used in infectious disease for identifying risk factors for disease occurrence, mortality and hospitalisation (Stoltzfus 2011). Both maximum likelihood and least squares are optimisation algorithms grounded in frequentist inference. Frequentist inference techniques are motivated by theoretical guarantees that apply if an identical data generating process could be observed independently many times. In contrast, Bayesian inference techniques are optimal with respect to the observed data in conjunction with pre-specified probability distributions capturing prior beliefs regarding the model and its parameters (Wagenmakers *et al.* 2008).

- **Bayesian approaches** obtain point estimates for parameters by updating a prior belief about the parameter using the likelihood calculated from the sample data to produce a posterior belief (Wagenmakers *et al.* 2008). Both priors and posteriors are expressed as probability distributions across all possible parameter values. Maximum a posteriori (MAP) estimates find the parameters that optimise the posterior possibility distribution, akin to maximum likelihood. The main difference is the likelihood is modified by prior beliefs. Prior distributions must therefore be carefully selected. In 'objective' Bayesian approaches, one may specify priors that reflect existing knowledge about the parameters (Robert 2007). When little

information is available, or its use deemed undesirable, an alternative objective approach is to specify 'vague' or 'uninformative' priors to reflect ignorance about the parameters (Robert 2007). In subjective Bayesian approaches, one instead tries to capture one's own personal beliefs about the parameters prior to observing the data. Regardless, the aim is to arrive at a joint posterior probability distribution over the parameters. This can prove computationally intractable depending upon model choice and the number of parameters contained, necessitating approximate numerical approaches. The Markov Chain Monte Carlo (MCMC) methods are a popular strategy for obtaining approximate samples from the posterior distribution (Gilks, Richardson, and Spiegelhalter 1995). MCMC iteratively explores the posterior distribution by performing a "random walk". During the walk, candidate estimates are successively drawn from a proposal distribution and accepted or rejected depending how much they improve the posterior distribution (Gilks, Richardson, and Spiegelhalter 1995). The posterior distribution is approximated by sub-sampling the chain at fixed intervals after discarding a fixed number of steps as burn-in (Gilks, Richardson, and Spiegelhalter 1995). From the approximated posterior distribution, a point estimate can be derived by several methods like posterior mean or median.

The parameters of most models can be defined with different estimators and under varying statistical doctrines. Logistic regressions used here as an example for maximum likelihood can also be fitted in a Bayesian framework (O'Brien and Dunson 2004). Further evading categorisation, regressions and their parameter optimisation are often restated as machine learning problems. Here, some parameters are typically estimated through cross-validation to minimise a loss chosen according to the dependent variable

(T. Hastie, Tibshirani, and Friedman 2008). The implementation of classical statistical tools such as logistic regression and ridge regression in machine learning blurs the dividing line between both disciplines. The cross-over of machine learning and statistical methods is unsurprising given that both aim to optimise parameters to fit a function to observed data. However, the goals of inference versus prediction may diverge beyond parameter optimisation. When machine learning has the stated goal of prediction, the point estimate of the parameter achieving optimal performance is typically sufficient for analysis. Whereas for statistical inference aimed at elucidating the underlying data generating process, quantifying the uncertainty of the estimates is of paramount importance.

1.6.2.2 Interval estimates

Point estimates for parameters may provide a concise summary of the model, but reliance on single numbers for describing data generating processes obfuscates their fundamental uncertainty (Demortier 2013). Defining parameter ranges to contain all estimates reaching some level of plausibility is a more faithful representation of inference concerning an unknowable process through random samples (Neyman 1937). As in point estimation, there are several methods to calculate parameter intervals that quantify uncertainty:

- **Standard errors** of a parameter estimate are equal to the standard deviation of its sampling distribution, a frequentist concept corresponding to the probability distribution of the estimate if the data generating process could be observed independently many times. The sampling distribution can be calculated theoretically in simple models, otherwise it must be estimated (Altman and Bland 2005). An

unknown standard error can be obtained by repeating parameter estimation for randomly resampled draws of the data, with replacement, in a process called bootstrapping (Efron and Tibshirani 1994). Intuitively, our confidence in estimating a population parameter increases the bigger our sample is, which is reflected in a decreasing standard error with increasing data quantity (Altman and Bland 2005).

- **Confidence intervals** define a range of parameter estimates wherein the 'true' population parameter lies with a quantified level of confidence, most commonly 95% (Campbell and Swinscow 2009). This is a frequentist concept, in which the distribution is constructed with respect to independent observations of the same data generating process. When faced with known population sampling distributions confidence intervals can be calculated directly. If the dependent variable is the result of a Gaussian distribution for example, the 95% confidence interval is the range of 1.96 standard errors above and below the point estimate. If the sampling distribution is unknown, bootstrapping can again be used. Confidence intervals are then achieved by defining a range of values that contains 95% of the parameter estimated from resampling (DiCiccio and Efron 1996). There are several alternatives to bootstrapping which require additional assumptions regarding the sampling distribution of the estimate, but can be more robust (Puth, Neuhäuser, and Ruxton 2015). Similar to the standard error, the width of the interval decreases with increasing data, assuming the variability does not increase (Hazra 2017).
- **Credible intervals** are the Bayesian counterpart to confidence intervals, giving a range of values wherein the true population estimate falls within a level of confidence (Eberly and Casella 2003). Credible intervals summarise uncertainty in

the posterior distribution: any parameter range that contains the required posterior probability is valid (Makowski, Ben-Shachar, and Lüdecke 2019). There are two common methods of defining credible intervals from the posterior. For unimodal posteriors, the highest posterior density interval chooses the narrowest interval which contains the required mass (Demortier 2013). The equal tailed interval chooses a range so that there is identical probability mass below and above the interval (Demortier 2013). As stated above, it can be computationally infeasible to evaluate the posterior for all possible parameter values with non-zero probability in the prior. Credible intervals are readily obtained using Monte Carlo sampling methods such as MCMC (Liu, Nordman, and Meeker 2016).

Parameter estimation describes systems empirically, by finding the most likely parameters of an underlying data generating process and quantifying the reliability of the estimates, whilst capturing the relationship between independent variables and dependent variables. Testing whether the estimated parameters are different from some pre-specified values, acting as a null hypothesis, is a closely connected technique for statistical inference, achieved through hypothesis testing (Koch 1999).

1.6.3 Hypothesis testing

Using knowledge to verify (or at least falsify) a theory through observation is central to the scientific endeavour and can be mathematically formulated as hypothesis testing (Guyatt *et al.* 1995). Theories formulated mathematically as hypotheses cannot be outright proven through empiricism, but statements of whether the data at hand rejects a given hypothesis can be made (Shi and Tao 2008). Hypothesis testing requires a definition of the null hypothesis is to be tested, alongside an alternative hypothesis

which is favoured upon rejection of the null (Guyatt *et al.* 1995). In the null hypothesis we condense our knowledge about the data generating process into a model and use the data to test whether the evidence supports this model. The null hypothesis usually states that no significant statistical relationship exists between a given independent and dependent variable, while its rejection indicates the existence of an association researchers are typically interested to discover (Haldar 2018).

There are two main errors that need to be controlled whilst testing hypotheses. A Type I error is the erroneous rejection of a genuinely true null hypothesis: a false positive. A Type II error is the failure to reject a false null hypothesis: a false negative. There is a trade-off between controlling false positives versus negatives. The typical view holds that false positives are more damaging than false negatives in the scientific process, perhaps because it is better not to take a step forward than to take a step backward, or perhaps because the process would otherwise be engulfed by noise.

In the frequentist approach, an acceptable type I error rate is pre-determined (Banerjee *et al.* 2009). The frequentist error *rate* is predominately defined conditional on the truth: the number of type I errors (false positive rate) is therefore quantified relative to the number of genuinely true hypotheses. (In contrast, one could condition on the inference: the false discovery rate (FDR) is the number of type I errors relative to the number of rejected null hypotheses.)

Commonly, frequentist hypothesis tests are performed subject to a type I error rate of 0.05, also called alpha level or significance threshold, so that the null hypothesis is mistakenly rejected in no more than 5% of the tests (Guyatt *et al.* 1995). If the test successfully rejects the null hypothesis at the defined alpha level, the association is

considered significant (Guyatt *et al.* 1995). With an increasing number of tests performed, the threshold can be adjusted to be more stringent using multiple testing correction methods like Bonferroni (Bonferroni 1936). In the case of GWAS examining multiple million individual mutations, an uncorrected error rate would lead to thousands of mutations falsely declared to be associated with disease, even if the null hypothesis was true. This would have catastrophic effects for downstream processes like drug and treatment design (Kaler and Purcell 2019).

Similar to the conditioning of an analysis on a type I error, the desired type II error rate can also be pre-determined in some settings, notably in the design of experiments. In clinical trials it is therefore common to pre-specify both the false positive rate and the power (also known as recall or sensitivity, see Figure 1-3) in order to determine the sample size (Lenth 2007). A power level of 0.8 is often chosen, from which the sample size required to detect a true effect from the sample can be calculated. Power calculations offer a benefit over machine learning, where insufficient data is also detrimental, but sufficiency can only be empirically defined after parameter estimation. Conditioning the analysis on specific type I and type II thresholds offers further benefits in statistical inference compared to machine learning. False positive and false negative rates can be pre-defined, and the parameter of the model can be estimated accordingly (Banerjee *et al.* 2009). In machine learning a threshold is defined either empirically or by optimising for a performance metric of choice after parameter estimation.

There are multiple methods of Bayesian and frequentist hypothesis testing with different approaches to determine significance thresholds while controlling type I and II errors:

- A common example of **Frequentist** hypothesis testing is the likelihood-ratio test that uses a maximum likelihood approach of finding which parameters best describe the data (Birkes 2005). The likelihood ratio test can only compare two nested models, where one hypothesis is a special case of the other, for example having one or more regression coefficients constrained to be 0 in the null hypothesis but not in the alternative hypothesis (Allen 1997).
- The test quantifies the difference of "goodness-of-fit" of both models by estimating the likelihood of the sample data given the two hypotheses (Birkes 2005). The difference is quantified by the test statistic, which is the log ratio of the likelihoods under the null versus alternative hypotheses (Birkes 2005). "Extremeness" is measured in terms of the test statistic, which allows for the definition of a p-value denoting the probability of observing data at least as extreme as the data observed if the null hypothesis were true (Birkes 2005). If the p-value is smaller the pre-defined type I error, the null hypothesis is rejected in favour of the alternative hypothesis (Woolf 1957). The inability to test two competing non-nested hypotheses is a limitation of likelihood ratio tests, which is not shared by Bayesian hypothesis testing (Pesaran 1990).
- In **Bayesian** hypothesis testing, two competing hypotheses are compared via the Bayes factor: the ratio of likelihoods under the null versus alternative hypotheses, akin to the frequentist likelihood ratio. The first difference relates to the handling of unknown parameters. In the Bayes factor, unknown parameters are averaged over by integrating with respect to their prior distributions (rather than plugging in the maximum likelihood estimates) (Baig 2020). The second difference relates to the

choice of threshold. The Bayes factor can be interpreted in terms of odds where a value of 3 indicates the data supports model in the numerator three times more than the model in the denominator (Hoijtink *et al.* 2019). While the Bayes factor quantifies the support of the data for one model over the other, it cannot prove either of them to be correct (Hoijtink *et al.* 2019). A threshold is typically pre-specified in terms of a large odds ratio (e.g. 20:1), directly, or via the posterior odds ratio which factors in the prior model odds. Unlike its frequentist analogue, the Bayes factor can be used to compare the support of the data for any two hypotheses, nested or not.

Hypothesis testing complements parameter estimation with its ability to test for associations between independent and dependent variables, which is invaluable for infectious disease research. The effect of medication, lifestyle or comorbidities on the risk of infection can readily be inferred within the hypothesis testing framework through, for example, risk factor analysis (Eckhardt *et al.* 2020).

As statistical models can uncover the association of independent values through inference, it would seem natural to assume inference can discover good predictors. However, this has not been observed in studies testing for significance, as has been shown in GWAS (Jakobsdottir *et al.* 2009; Clayton 2009; Janssens and van Duijn 2008), the reason for which also lies at the heart of the prediction and inference dichotomy: confounders. As confounding variables influence both independent and dependent variables, a non-causal association between independent variables and the dependent variable is created. The relationship is non-causal as a change in the observed independent variable would not necessarily lead to a change in the dependent variable.

A change in the confounding variable however, would lead to changes in both the associated independent variable and the dependent variable. Statistical inference is concerned with gaining insight of an unknowable population through a sample. Therefore, wrongly assuming associations caused by interactions with variables outside the sample to be causal, is an egregious mistake in statistical inference (Greenland, Pearl, and Robins 1999). This can be addressed in the statistical framework by using one of multiple methods designed explicitly for controlling confounders (see Brenner and Blettner 1997, for example). When the goal is to predict dependent variables from independent variables in machine learning, we are not interested in discovering the nature of the relationship so long as we can reliably predict outcomes beyond training data. To illustrate, in the inference setting of GWAS, researchers aim to discover mutations that are causal for a disease to aid drug development among other goals. Discovering genetic confounders, such as markers of ethnicity which might indicate an infection being spread in a specific community, would not be helpful for drug development (Hu and Ziv 2008). In a genetic machine learning setting, where we are trying to assign disease presence or absence from genomes, markers of ethnicity can improve the accuracy of the prediction if the disease is more prevalent in some ethnic groups. When controlling confounders, focussing only on significant interactions can decrease the predictive information contained in all possible interactions between independent and dependent variables. It seems the trade-off described by the “no free lunch theorem” again holds true: methods designed for inference are not equally proficient in uncovering predictors.

1.7 Statistics Versus Machine learning: Right Tool for The Right Job

As demonstrated above machine learning and statistical inference serve different purposes, resulting in different constrains on analysis which manifest in differing toolkits. Statistical inference further divides into frequentist and Bayesian frameworks, bringing further divide upon how to best gain knowledge from the data at hand. Some methods, however, are motivated on frequentist and Bayesian grounds whilst also being used in machine learning, showing how the distinctions I described above can be blurred. Ridge frequentist regression is one example which estimates parameters of multiple regression models (Mallick and Yi 2013). Ridge regression combats problems including over-parameterisation and multicollinearity by introducing a shrinkage parameter for regularisation (Hoerl and Kennard 1970). The shrinkage parameter can be estimated in a frequentist or Bayesian manner, or via empirical cross validation as a machine learner (Mallick and Yi 2013). The individual regression coefficients are estimated essentially via a prior distribution specified by the shrinkage parameter, which is known as a random effects model in the frequentist framework. The shrinkage offered by ridge regression is particularly important for big data settings with myriads of correlated independent variables, such as when using genetic data. GWAS determines the effect of single nucleotide polymorphism (SNPs) as independent variables with a phenotype of interest as a dependent variable, such as COVID-19 severity (The Severe COVID-19-19 GWAS Group 2020). Modern GWAS approaches typically estimate regression coefficients for every SNP; with two human genomes typically differing in 4 to 5 million sites (Auton *et al.* 2015) and studies rapidly approaching the 1 million participant mark (Hautakangas *et al.* 2022). Ridge regression

serves as a null model for GWAS, used as a baseline to test against the unconstrained association of individual SNPs with the phenotype, over and above the influence of the regularisation constraints imposed by the ridge regression (J. Zhang *et al.* 2021). This is known as a linear mixed model (LMM). If the difference in model performance between null model and the model with the addition of a single SNP with unconstrained effect size is significant, then an association is inferred (Appasani 2016). The parameter estimate of the regression coefficient gives the directionality and magnitude of effect of the mutation (de Vlaming and Groenen 2015), while the underlying ridge regression serves as a method of controlling confounders, both genetic and non-genetic (A. L. Price et al. 2010). In the machine learning framework ridge regression is used for prediction where the regularisation is introduced to prevent over-fitting (Rhys 2020), to select input features (Shichao Zhang et al. 2018), or to adjust output weights in artificial neural networks (G. Li and Niu 2013). Ridge regression demonstrates how the same tool can be used for varying means whilst also adhering to different philosophies.

Even if used under multiple doctrines, generating insight in the big data era of infectious disease has not been accomplished by a singular tool or single collection of related methods. With the apparent performance trade-off between prediction and inference being expressed in differences in methodology, big data analysis requires the right tool for the right job. Both prediction, e.g. mortality prediction, and inference, e.g. risk factor analysis, are important for understanding infections, leading to an abundance of varying methods deployed in the field. My motivation in the following was not to explore all applications of a single approach, but to use methods taken from both machine learning and statistics to tackle current problems in infection using big data:

- As campylobacteriosis is the most prevalent food-borne disease, attributing the source of infection is an important tool in combatting outbreaks and the administration of public health interventions. Currently the most prevalent source attribution technique does not reflect the state-of-the-art data types that are available in the age of next generation sequencing. Including more complete captures of the bacterial genome could improve accuracy and would enable the assignment of sources to individual isolates, which is currently impossible. My aim in **Chapter 2** is to address this shortcoming by reframing source attribution as a machine learning prediction task. I apply gradient boosted trees to attribute the source of Campylobacteriosis infections in humans based on genetic differentiation between reservoir populations. I report a 33% increase of accuracy over the currently most widely used attribution techniques and enable prediction of source for individual samples which showed a change in host affinity in one of the most prevalent lineages in human infection.
- Although *C. jejuni* is the most common cause of gastroenteritis, the understanding of its pathogenicity lack behind less prevalent microbes like *Escherichia coli*. Inferring knowledge about the different routes of infection from varying niches is further hindered by individual attribution of human samples being impossible. In **Chapter 3** I use the source attribution I developed to fill the knowledge gap by using the prediction as starting points for statistical inference. I use a GWAS to identify genetic variation that differentiates *Campylobacter jejuni* isolates from different host species using whole genome sequences. I show that pathways involved in surviving the microenvironments of poultry production and intestine to be significantly associated with host specificity.
- With the instantaneous urgent demand of COVID-19 related insight, multiple studies have successfully leveraged the data contained in large well-curated databases to

perform risk factor analysis for disease severity. However, as complex diseases like COVID-19 are the result of intricate and interacting pathways, the risk factor discovery is heavily influenced by the prior choice of which variables to include. As there is currently no principled way for agnostic risk factor prediction, I aim to provide one alongside a prediction of COVID-19 severity risk in **Chapter 4** which considers both prediction and inference problems. Using gradient boosted trees, I predict susceptibility to SARS-CoV-2 and the risk of different disease outcomes based on UK Biobank data. To subsequently assess risk factors, I use Bayesian point estimation to fit a logistic regression to a vast array of available measurements contained in the UK Biobank. As machine learning is a convenient tool for analysis large heterogeneous datasets, it is sometimes also used to interrogate risk factors, which is problematic as it does not address confounders. To show the shortcomings of this approach, I compare different measures of feature importance in machine learning to the posterior inclusion probability of the Bayesian analysis. My comparisons show that importance assigned to individual variables by algorithms optimised for prediction, differs from the relative importance estimated by Bayesian inference. My findings further show that socio-demographic factors and descriptors of ill health are important for disease outcome alongside history of prior lung injury.

Apart from the domain specific knowledge generated in my work, I explore how to leverage the respective strengths of machine learning and statistical inference and how they can be complementary. Despite their differing motivations and data philosophies, the synthesis of machine learning and statistics holds tremendous power at the dawn of the big data era in infectious disease.

Chapter II

Machine Learning to Predict the Source of Campylobacteriosis Using Whole Genomes

Overview

Campylobacter jejuni is the main aetiological agent of the world's most frequent foodborne illness, Campylobacteriosis. As there is a broad spectrum of origins of infections ranging from contaminated meat and poultry to drinking water, the designing of interventions hinges on accurate source attribution. The varying host affinities of different strains and their genetic manifestations can be leveraged for source determination through multi-locus sequence typing (MLST) schemes. Human isolates of Campylobacteriosis cases can thus be assigned sources based on the allelic variation observed in the source samples. The limited genetic capture of the seven MLST genes has bounded the maximum achievable accuracy of the probabilistic methods based on the sequencing scheme invented in the late 90s. My work broadens the spectrum of possible inputs to include core genome MLST (cgMLST) and whole genome sequences (WGS), whilst also improving the maximum achievable accuracy by 33% with the best performing classifier amongst multiple machine learning algorithms I trained. The previously most frequently used iSource method resulted in an attribution accuracy of 64%, whereas on the same data the top performing machine learner achieved 71%. The cgMLST resulted in 85%, and the WGS approach was 78% accurate. Predicted sources show the unique ability of chicken niche specialists to infect humans, whereas generalists from other sources are the major contributors to human disease. To leverage the increased accuracy for furthering the understanding of Campylobacteriosis, a Bayesian inference algorithm was used to investigate the relative infectivity of *C. jejuni* strains within ST-21. Alongside changes in transmissibility to humans, I also identified differences in host affinities within the closely related isolates. By providing an easy-to-

use classifier that can attribute individual isolates with high generalisability and low computational requirements I offer a scalable approach to the global surveillance of Campylobacteriosis, which can be readily improved when given new data.

2.1 Introduction

Gastroenteritis is responsible for an estimated nine million cases in the European Union annually and is caused primarily by the bacteria *Campylobacter jejuni* and *Campylobacter coli.* ('The European Union One Health 2018 Zoonoses Report' 2019; Kaakoush *et al.* 2015). *Campylobacter* bacteria are zoonotic microbes commonly found in the gut of birds and other animal species as a commensal (Sheppard et al. 2011a; Sheppard, Colles, et al. 2010a) but can cause serious infections when transmitted to humans. Campylobacteriosis is commonly accompanied by symptoms including nausea, fever, abdominal pain, and severe diarrhoea, with a low chance of causing debilitating, and sometimes fatal, sequelae (Nachamkin, Allos, and Ho 1998; Nielsen *et al.* 2010). Sources of infection can be animal faeces, contaminated drinking water and, most frequently, raw or under-cooked poultry and chicken or sheep meat (Altekruse *et al.* 1999). Administering effective interventions requires detailed understanding of the relative contribution of different sources to human infection.

As is common in bacteria, *Campylobacter* populations consist of diverse assemblages of strains (Sheppard *et al.* 2011a; Gilbert *et al.* 2016; Kirk *et al.* 2018; Sheppard, Dallas, *et al.* 2010). Some lineages within the structured population exhibit an increased host affinity towards one or multiple source animals, which can be leveraged for host attribution (Sheppard et al. 2011a; Sheppard, Colles, et al. 2010a; Ogden et al. 2009). Comparing the genetic variability of human isolates with *C. jejuni* samples from all

potential reservoir species allows for a probabilistic assignment of host origin based on observed similarity.

Past source attribution efforts have determined raw or undercooked chicken meat as the main contributor to transmission to humans (Institute of Environmental Science and Research Ltd 2007; Sheppard, Dallas, MacRae, et al. 2009a). The development of targeted intervention was enabled by genetic source attribution (Nichols *et al.* 2012), most notably improved biosecurity measures on chicken farms halving disease incidence in New Zealand in 2007 (Sears *et al.* 2011; Nohra *et al.* 2020).

The concept of connecting source to sink populations by genetic similarity was used in the development of methods assigning likely sources of human isolates based on allele frequencies observed in host reservoirs (Wilson et al. 2008a; Sheppard, Dallas, Strachan, et al. 2009a). The earliest methods are based on multi-locus sequence typing (MLST), one of the most abundant genotyping data for *C. jejuni* which is still the basis for most source attributions (Cody *et al.* 2019). MLST captures the genetic variation observed in seven housekeeping genes that are commonly found in all strains (Maiden et al. 1998a; Dingle et al. 2001). Samples are designated as the same sequence type (ST) if all seven alleles are identical, and as clonal complexes (CC) group if at least five loci are the same. The combination of observed allele frequencies in animals with genetic groupings allowed for the probabilistic assignment of the most likely source origin of STs and CCs. The main models for attribution are the asymmetric island model implemented in iSource (Wilson et al. 2008a) and the Bayesian population assignment model STRUCTURE (Pritchard, Stephens, and Donnelly 2000; Sheppard, Dallas, Strachan, et al. 2009a). The application of both approaches has allowed for the determination of the

relative contributions of several farm and wild animal hosts to human transmission, with poultry being the foremost source of human gastroenteritis caused by *C. jejuni* across different regions and countries (Wilson et al. 2008a; Sheppard, Dallas, Strachan, et al. 2009a; Mullner et al. 2009; Boysen et al. 2014; Di Giannatale et al. 2016; Kittl et al. 2013).

There are two core caveats of using bacterial genetic variation for source attribution. The first is that attribution ability is limited by the degree of genetic separation recorded in the underlying data. Genotypes in *C. jejuni* can be restricted to hosts (Sheppard *et al.* 2011a; Mourkas *et al.* 2020), allowing for accurate attribution. However host niche generalists, which have relatively recently switched hosts, also exist (Sheppard, Cheng, Méric, Haan, et al. 2014; Woodcock et al. 2017), hindering the assignment of a definite source (Dearlove et al. 2016a). Host switching blurs the genetic separation leveraged by methods of quantitative attribution resulting in lower accuracy which is difficult to circumvent. The second caveat of host attribution can be more easily remedied. Specifically, the current most frequently used source attribution models are limited by the data type they have been developed for. Being a product of the state of technological progress in the late 90s, MLST only captures a miniscule part of the genome (~ 0.2% in *C. jejuni* (Kittl *et al.* 2013)), where modern sequencing technologies offer far greater potential at differentiating lineages but are currently not used for source attribution.

In the following, I develop a machine learning approach that leverages current genotyping technology by using WGS data for *C. jejuni* source attribution of human isolates. This offers two advantages. Firstly, an agnostic approach to algorithm choice

where a wide array of general-purpose algorithms is included. This was proven successful by WGS-based machine learning source attribution approaches in *Salmonella enterica* and *Escherichia coli* (Shaokang Zhang et al. 2019; Lupolova et al. 2017). Secondly, I utilise several different WGS data types to facilitate the analysis of existing MLST, core-genome (cg)MLST and WGS datasets and reuse of data for streamlining genomic disease surveillance. My goal is to overcome existing limitations in source attribution methodology and demonstrate the ability of the resulting algorithm to generate insight into the differing infectivity of closely related *C. jejuni* lineages.

2.2 Methods

2.2.1 Dataset Collection and Preparation

A total of 5,798 *C. jejuni* and *C. coli* genomes isolated from various sources and host species were available on the public database for molecular typing and microbial genome diversity: PubMLST (https://pubmlst.org/). WGS data corresponded to MLST ST and CC designations as well as cgMLST classes. The dataset was divided into training (75%) and testing (25%) sets using phylogeny-aware sorting, wherein all members of one ST were sorted entirely into either training or testing sets. The ST based sorting accounts for the phylogenetic non-independence of samples (Lees *et al.* 2020). To allow for sufficient sample sizes per reservoir population (hereafter "class"), only the five most prevalent classes for MLST and cgMLST were used (chicken, cattle, sheep, wild bird, and environment). For farm animals the classes "chicken" and "chicken offal or meat" were combined to "chicken" (likewise for sheep and cattle), whilst "environment", "sand" and "river water" were combined into "environment", consistent with previous studies (Sheppard, Dallas, Strachan, et al. 2009a; Thépault et al. 2017).

2.2.2 Feature Engineering

The allelic profiles of MLST and cgMLST were used directly. To exploit the gradient of separation encoded in the sequences underlying the MLST allelic profiles, allele sequences of the loci were downloaded and nucleotides encoded as dummy variables and k-mers (k=21) using DSK (Rizk, Lavenier, and Chikhi 2013a). DSK was also used for encoding the WGS as k-mers. Using k=21 led to a prohibitively large input vector due to the number of unique k-mers found in all genomes (109,675,176). The number of k-mers was reduced by applying a variance threshold where k-mers that were present or absent in more than 99% of the samples were discarded, reducing the numbers of unique k-mers to 7,285,583. Furthermore, feature selection was performed by testing the dependence of the source labels on every individual k-mer using the Chi-Square statistic. To avoid data-leakage the feature selection was used before data was split into training and testing to select the 100,000 k-mers with the highest score.

2.2.3 Algorithm Training

All machine learning and deep learning was performed in Python (for a list of all algorithms see Figure 2-1). The XGBoost library (T. Chen and Guestrin 2016) was used for the gradient boosting classifiers with all other machine learners implemented in scikit-learn (Pedregosa *et al.* 2011). The hyper-parameters for each classifier were chosen using Cartesian grid search on five-fold cross-validation of the training set. The Keras library (https://github.com/keras-team/keras) was used to construct deep learning algorithms aimed at supplying a wide range of commonly used architectures. This was empirically found to be best, given that there is no standardised methodology for architecture selection of such models. Specifically the following were used: (i) A

recurrent neural network consisting of a layer with 64 gated recurrent units, a 50% dropout layer and Rectified Linear Unit (ReLU) activation layer; (ii) A 1-dimensional convolutional network with two convolutional layers of kernel size 3 and 5 respectively and 30 filters, both followed by 50% dropout layers and a ReLU layer; (iii) A Long short-term memory (LSTM) network consisting of one LSTM layer with 64 units and a 50% dropout layer; (iv) A Shallow dense network with one dense layer with 64 units followed by a 50% dropout layer and a ReLU activation layer; (v) A Deep dense network with 6 dense layers starting with 128 units and halving units with each successive layer. All individual dense layers are followed by a 50% dropout layer and a ReLU layer.

To all deep learning architectures, an output layer comprising a dense layer with soft-max activation with one unit for every class was added. The labels were encoded as dummy variables and categorical cross-entropy was used as a loss function together with the Adam optimiser (Kingma and Ba 2014). Cyclical learning rates were used with a maximum learning rate of 0.1 and a minimum learning rate of 0.0001 to overcome local minima. The accuracy on the test set was measured at every epoch and the overall best performing weights were stored as a checkpoint. The data was deployed in batches of 128 samples with every batch randomly undersampled so that each class was represented in equal proportions. The training was run for 500 generations with early stopping after 50 generations.

2.2.4 Algorithm Testing

Both machine learning and deep learning were tested on the same 25% test set. The original data were skewed in source composition by ratios which did not necessarily

reflect source origin of infection. Therefore, two methods were used to rebalance the classes in testing. The first test set featured an even distribution of classes, whereas the second undersampled the over-abundant chicken-origin genomes to emulate relative contribution to human disease. The ratios predicted by Wilson *et al.* (2008) were used, where *Campylobacter* genomes from chickens were 1.61 times more common than those from cattle. In both methods, rebalancing the classes was achieved by undersampling, which I repeated 200 times with replacement and averaged the accuracy over all iterations whilst also recording the variance. Accuracy, precision (positive predictive value), recall (sensitivity), F1, negative predictive value, specificity and speed were recorded as performance metrics (see Figure 1-3 in the introduction for an explanation of performance metrics). Speed was measured relative to other classifiers and a scale was defined with 0 being the slowest classifier and 1 being the quickest and all intermediate values being normalised within these confines. For comparison to previous methods, iSource was applied to the test dataset (Wilson et al. 2008a). Having established that XGBoost on cgMLST was the best performing source attribution method, the classifier was retrained with both training and testing data and applied it to all 15,988 human cgMLST samples available on the PubMLST database. The prediction took 892 milliseconds on a Dell OptiPlex 7060 desktop using ten threads on an Intel Core i7-8700 CPU and 16 GB RAM.

2.2.5 Phylogenetic Analysis

The generalist index was defined as the number of sources the ST was found in across all isolates in the dataset, which included additional samples for which only MLST data was available. A phylogeny of CC21 genomes from both source-associated and human

isolates was built using Neighbour Joining, based on pairwise hamming distances of k-mer presence/absence in the WGS dataset, as described by Hedge and Wilson (Hedge and Wilson 2014). TreeBreaker was used to infer the evolution of phenotypes across the phylogenetic tree of ST-21 and the most closely related sequence types. The known labels of the source-associated samples were used as phenotypic information for input into TreeBreaker (Ansari and Didelot 2016) together with the phylogeny of CC21. TreeBreaker was run for 5,500,000 iterations with 500,000 iterations as burn-in and 1000 iterations between samplings. The phylogenetic trees were visualised with Microreact (Argimón et al. 2016a) and arranged alongside the results of TreeBreaker in Inkscape.

2.3 Results and Discussion

2.3.1 Machine learning Outperforms Popular Attribution Models for MLST Data

To embed the source attribution accuracy in the performance of previous efforts, I used the asymmetric island model on the MLST dataset implemented in iSource, currently the most popular source attribution method (Cody *et al.* 2019). As seen in Figure 2-1, the random forest model which performed best on the MLST data (61.9%/68.5% balanced/unbalanced) performed slightly better than iSource (61%/64%) on the same dataset. Since allelic profiles do not consider the underlying sequences and only offer a binary comparison, matching or not, I incorporated allele nucleotides to explore whether the gradient of genetic differentiation offers better performance. I dummy encoded the loci sequences and used k-mers generated from the loci which boosted the best performing MLST classifiers to 67.9%/70.7% from nucleotide dummies and 63%/67.5% from k-mers.

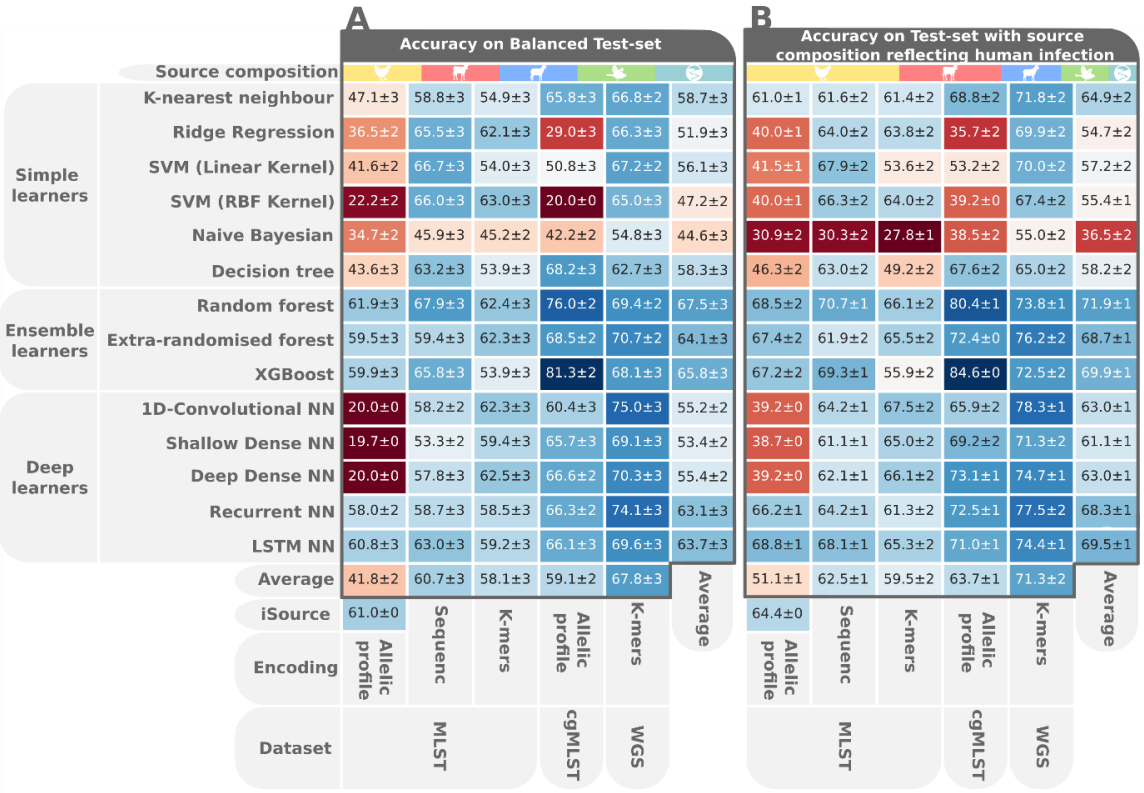


Figure 2-1 Performance of all classifiers

A heatmap showing classifier performance on the class balanced (A) and imbalanced (B) test set. The individual cells are coloured according to the average accuracy on 200 rounds of resampling with replacement with the variance noted next to the average accuracy. The averages of accuracy per classifiers are shown in the rightmost column, whereas the bottom column shows the averages per data type.

2.3.2 Core Genome and WGS Datasets Increase the Power Of Source Attribution Models

Having ensured the competitive performance of my machine learning algorithms to existing methods using MLST data, I focused on whole *C. jejuni* genomes. As the logical extension of seven-gene MLST, gene-by-gene approaches of genomic capture have been used in *Campylobacter* to transcribe a larger portion of the genome *in silico*. To describe the core genome, the cgMLST scheme was introduced that extended MLST to

1,343 loci present in the majority (>95%) of *C. jejuni* genomes (Cody et al. 2017a). CgMLST has the potential to offer greater discrimination of *C. jejuni* lineages which could improve source attribution efforts utilising genetic variation differing between hosts (Thépault *et al.* 2017). Tree-based ensemble classifiers continued their strong prior performance on cgMLST, with the XGBoost classifier achieving 81.3%/84.6% accuracy, which is also the highest accuracy over all data types and classifiers.

Next, I assessed the relative performance of machine learners when applied to k-mers produced from WGS, where the average attribution performance was the highest among all datasets. The best-performing algorithm was a 1-D convolutional neural net (75.0/78.3%), performing better than the top-achieving classifier on MLST but worse than the best classifier on cgMLST despite WGS encoding more genomic information. This may be explained by the feature selection used to limit the input vector to 100,000 k-mers.

Beyond comparing classifier performance on different data types, I also wanted to investigate what led to the difference in performance. Comparing the predictive performance throughout all data types shows increases in encoded variation led to an improved average accuracy across all algorithms. Specifically exposing additional gradual separation of the underlying sequences resulted in better performance in MLST, even with the same ~0.2 percent of the genome being encoded. The accuracy comparison of all classifiers reveals strong performance of decision-trees ensembles across all data types, with random forests being the best among all ensemble learners on average. Algorithms based on decision-tree ensembles have been reported to perform well given genomic input (Austerlitz et al. 2009; Deneke, Rentzsch, and Renard

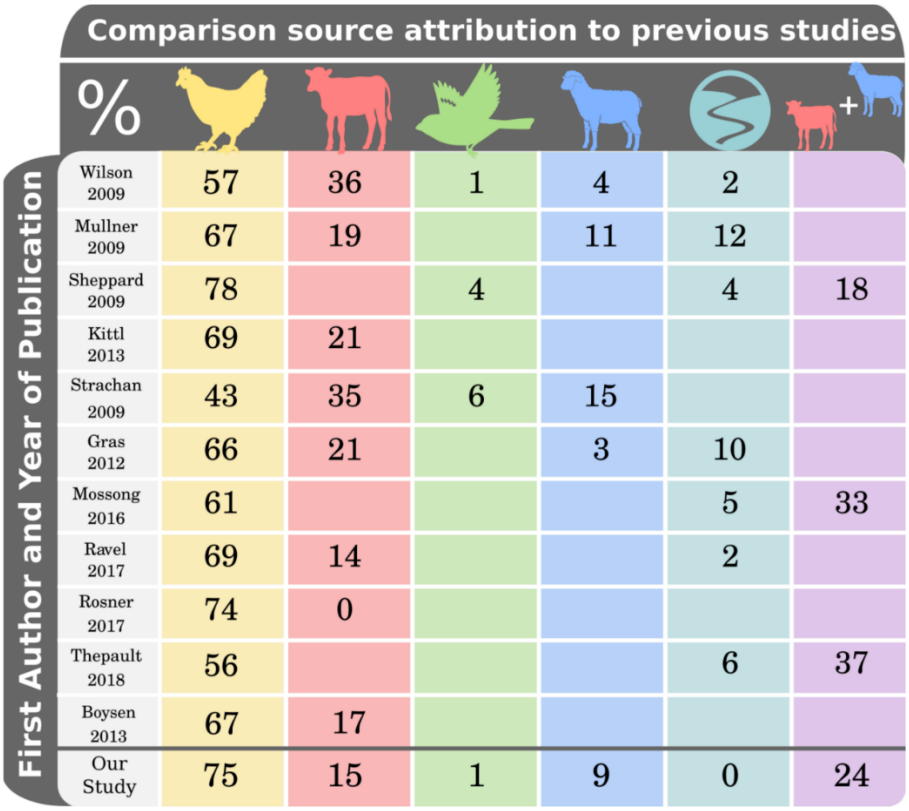
2017b; X. Chen and Ishwaran 2012), which has been attributed to their capacity to handle the interaction and correlation of features innate to genomic data (X. Chen and Ishwaran 2012).

The highest performing simple learner was the k-nearest neighbour algorithm (KNN), which is potentially the result of the phenotypic trait to be predicted being hereditary. The ability to colonise specific hosts is passed down genetically with environmental factors also contributing to passing on the trait, as the proximity between parent and offspring cells is required. The mechanisms of inheritance lead to more closely related sequences having a higher probability of being associated with the same phenotype. Heritability could be causal for the observed success of the KNN algorithm. KNN assigns class based on neighbourhood in the hyperdimensional feature space (Kotsiantis, Zaharakis, and Pintelas 2006). The feature space in my captures work genetic similarity which also captures relatedness.

The deep learning algorithms used here generally showed higher average accuracy with increasing dimensionality of the input data from MLST to WGS. Amid all deep learners, the RNN and LSTM achieved the highest accuracy, which was unsurprising given DNA is transcribed, and mRNA translated, successively from 5’ to 3’ end. This biological process is mirrored in RNNs and LSTMs, as both process input data sequentially and input weights are also tuned successively during back-propagation. The weights of the less well performing convolutional neural nets for example are fine-tuned concurrently. After the comparison of throughout all algorithms and data types, I wanted to assess the source attribution of the best-achieving classifier against the previous source attributions.

2.3.3 Machine learning Source Attribution is on the High End of Previously Reported Chicken Attribution Bound

The overall highest accuracy across all datasets was achieved by gradient boosted decision tree using the XGBoost implementation on the cgMLST dataset, which was used for a more thorough analysis of performance. My source attribution, which I termed aiSource, compared to previous source attribution efforts can be seen in Figure 2-2-2, which reveals that my assignment of human cases to poultry is close to the higher bound of previous poultry attributions. Predicting more sources to come from chicken also results in lower assigned cases to come from all other niches, with only 0.4% of human cases predicted to have environmental origins. The diverging source attribution of aiSource to previous efforts could be the result of the greater discriminatory power the cgMLST data holds compared to MLST. To lend additional credibility to my source attribution beyond the comparison to other studies, I investigated the influence of possible confounders on source attribution accuracy.



First Author and Year of Publication	%					
Wilson 2009	57	36	1	4	2	
Mullner 2009	67	19		11	12	
Sheppard 2009	78		4		4	18
Kittl 2013	69	21				
Strachan 2009	43	35	6	15		
Gras 2012	66	21		3	10	
Mossong 2016	61				5	33
Ravel 2017	69	14			2	
Rosner 2017	74	0				
Thepault 2018	56				6	37
Boysen 2013	67	17				
Our Study	75	15	1	9	0	24

Figure 2-2-2*Comparison of my source attribution to previously published studies*

2.3.4 Host transition Imposes a Biological Limit on Source Attribution Models

To understand the limitations of aiSource, I stratified error rates into factors potentially driving misclassifications in the different datasets as seen in Figure 2-3. The classifier most frequently mistakes isolates from sheep and cattle for each other, which is also a common error of other methods of source attribution (Wilson et al. 2008a). Frequent host switching resulting in overlapping gene pools could explain the misclassification and suggests a similarity in the ruminant gastrointestinal tract's physiology (Kwan *et al.* 2008). Beside host reservoirs, I further investigated other sources of error, as the season in which the sampling occurred (Méric *et al.* 2018) and geographical origin of the isolate (Kwan *et al.* 2008) are known confounders of source attribution. To this end, I stratified attribution performance by *Campylobacter* species, continent, year, and generalist index based on the full non-undersampled test dataset (Figure 2-4).

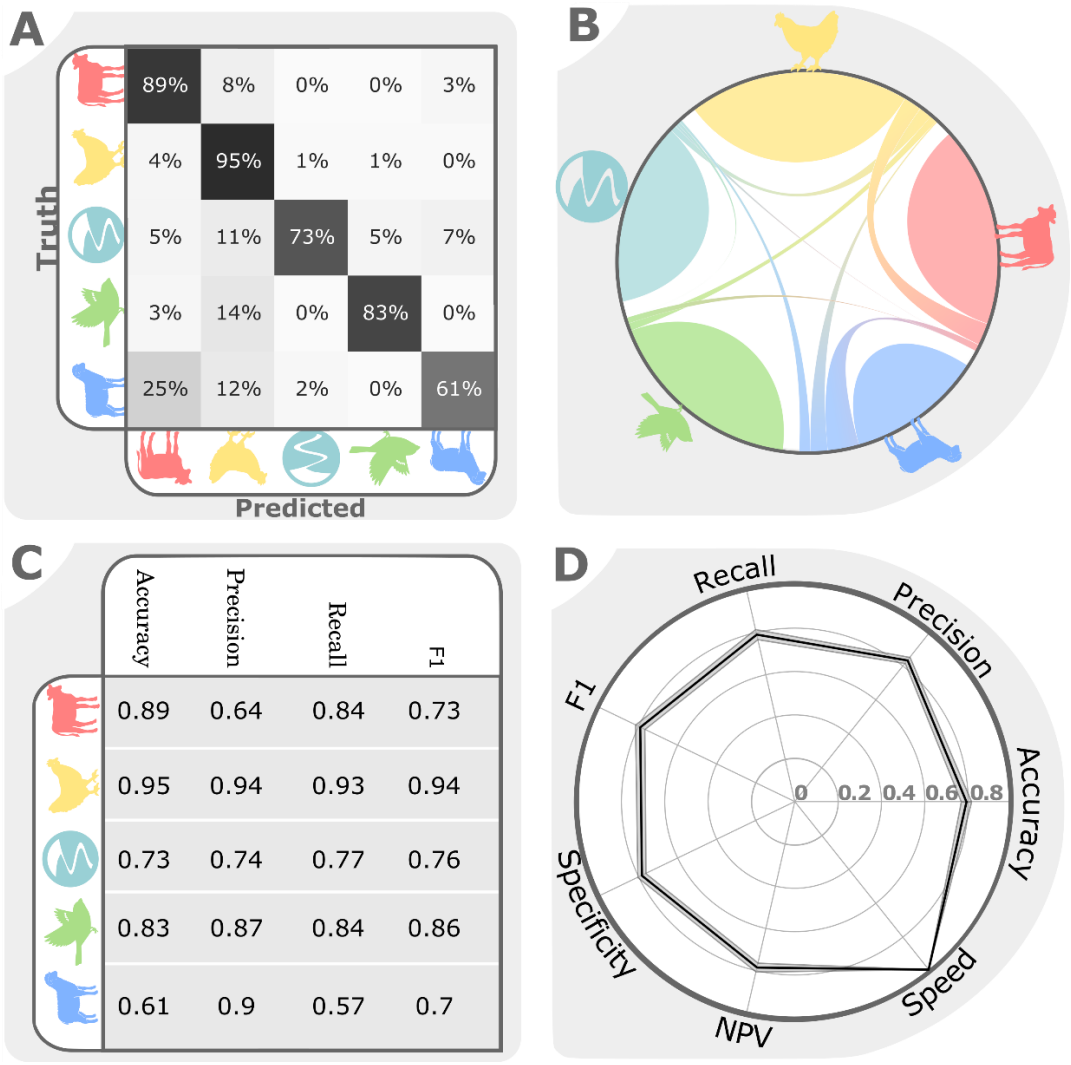


Figure 2-3 XGBoost Classifier performance on cgMLST

A) Misclassification matrix per source. The diagonal represents correct classification and off-diagonal fields are misclassifications. The percentages are calculated per row. B) Misclassification matrix as depicted in a flow diagram. C) Classifier performance on the unbalanced test set according to four different metrics per source population. D) Radar plot showing the classifier performance on the unbalanced test by seven metrics averaged over 200 rounds of resampling with replacement. The variation is depicted as a shaded surface underneath the black line representing the average.

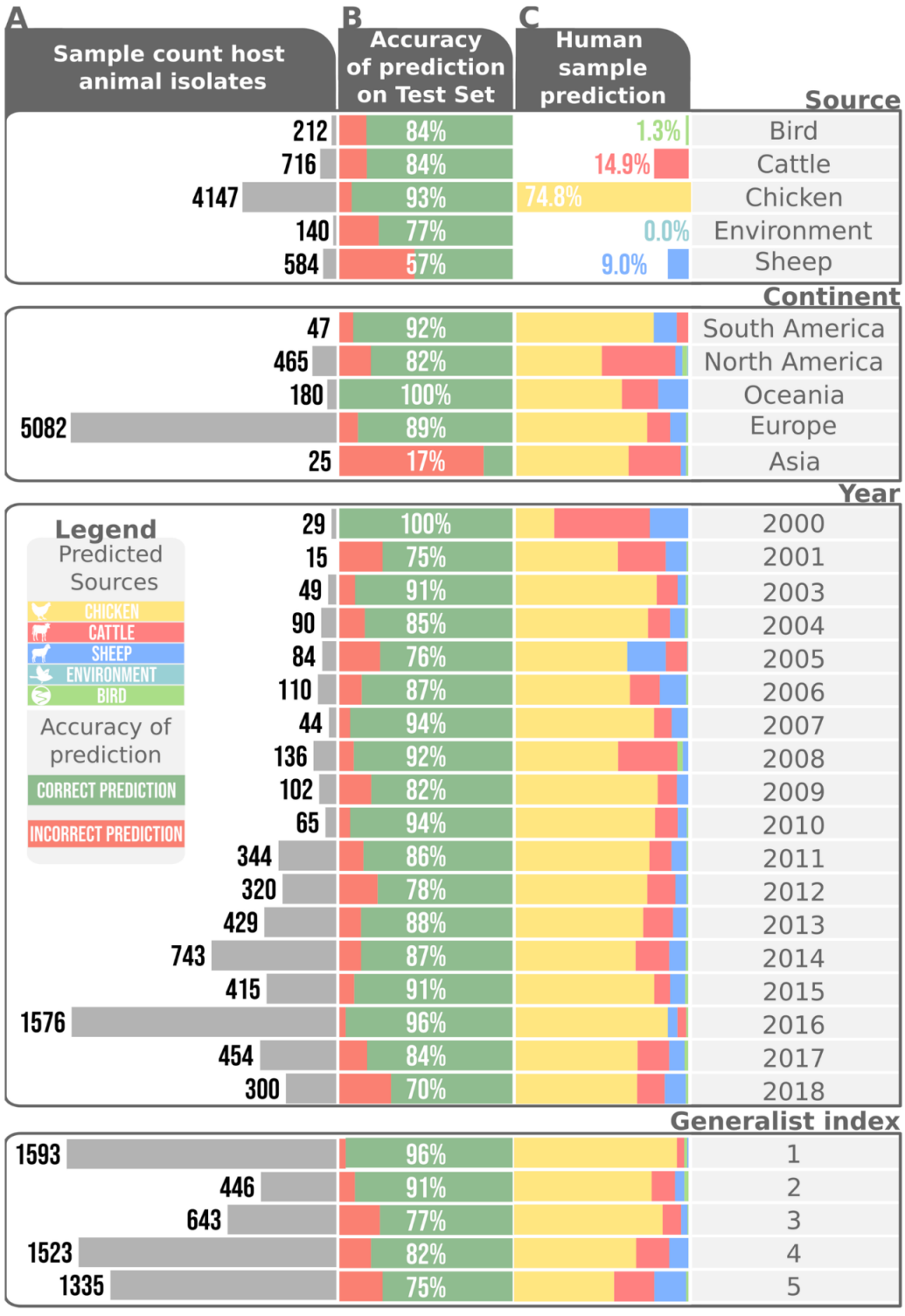


Figure 2-4 Stratification of source attribution.

Source attribution per source, continent, year generalist index and Campylobacter species. A) Sample sizes across different factors in the imbalanced training set. B)

Prediction accuracy on the full test dataset divided by different factors. C) Source attribution stratified into varying factors.

Approximately 10% of human Campylobacteriosis is caused by *Campylobacter* coli (Sheppard and Maiden 2015), another member of the *Campylobacter* genus. As it was readily available from PubMLST I wanted to explore whether my improved classification performance observed in *C. jejuni* extends to the closely related species. I found better attribution performance in *C. coli*, which is corroborated by previous investigations and possibly a result of the more strongly segregating lineages between *C. coli* hosts (Dearlove et al. 2016a). A higher relative contribution to human infection by isolates from sheep and the environment could further facilitate accurate attribution (Ogden *et al.* 2009; Roux *et al.* 2013; Strachan *et al.* 2009).

A closer inspection of the influence of the sample size on XGBoost performance revealed that the paucity of wild bird samples (212 samples; 84% accuracy) did not encumber classification performance relative to the more ample samplings of cattle (716 samples; 84% accuracy) and sheep (584 samples; 57% accuracy). Bird samples could be easier to classify due to lineages within wild birds being highly dissimilar from their farm counterparts, and cross-over to other host niches occurring only sporadically (Sheppard, Jolley, and Maiden 2012). It has been proposed that the ability to switch hosts can negatively impact attribution accuracy due to blurring the source-specific genetic fingerprint (Dearlove et al. 2016a). I investigated the impact of host switching on accuracy by defining a “generalist index” as the number of hosts in which an ST was found across all PubMLST samples. As expected, “specialist” strains isolated from fewer host animals were classified with increased accuracy. In concordance with my previous

findings 58% of all wild bird samples belonged to STs only found in this niche, whereas other environments exhibited lower proportions of specialists (environment = 41%, cattle = 9%, sheep = 3%, chicken = 32%). Despite varying across different generalist indices, performance was still adequate (lowest 75%) on the complete spectrum from specialist to generalist strains, which encouraged me to investigate how host switching influences infectivity based on my prediction.

2.3.5 Human Infections from Non-Farm Sources Are Primarily Caused by Generalist *C. jejuni*

A key selective force within *C. jejuni's* evolutionary trajectory is the adaptive bottleneck occurring during transmission, as not all lineages transmit from source animal to human. To delineate the evolutionary forces during the bottleneck I separated isolates into host reservoirs and compared source samples to human samples attributed to the same source (Figure 2-5). Using this comparison, I caught differences between the populations strain composition as captured in CC distribution before and after transmission. The frequency shifts in CC compositions appear larger for the environment and bird niche than for farm animals suggesting a narrower selective bottleneck, which is expected given lower transmission from non-farm sources.

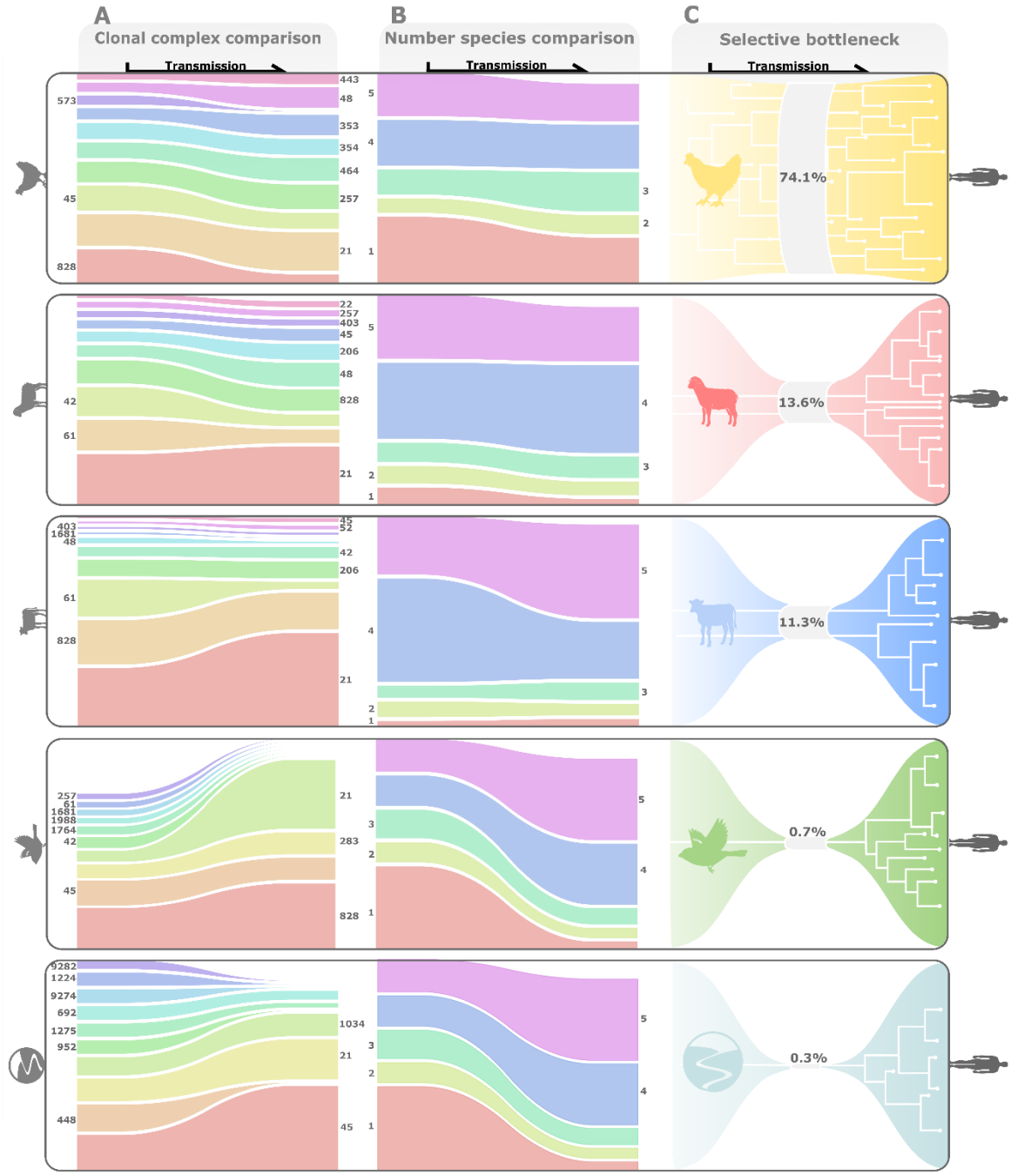


Figure 2-5 The adaptive bottleneck of C. jejuni transmission

A visual representation of the adaptive bottleneck occurring during zoonosis. A) Changes in frequencies between clonal complexes in isolates from sources directly and human isolates predicted to stem from the same source are shown. B) Changes in frequency of the number of sources the clonal complexes can be found in. Some lineages only appear

in humans and thus cannot be assigned a generalist index, which is expressed in the white space on human side of transmission. C) Size of the selective bottleneck is visualised through the proportion of human infection. The phylogenetic trees shown within the bottlenecks are for illustrative purposes only.

We also calculated the relative composition in generalist index before and after transmission from the CC composition (Figure 2-5B). Through analysing the shift in the generalist index, I gained insight of how the ability to colonise multiple hosts affects human infection specifically. In the non-farm niches the shift in generalist index shows that numerous CCs unique to environment and bird (and to a lesser degree ruminants) are either absent or less frequent in human samples. The frequency of generalists rises after transmission in their stead which suggests infectious lineages from the non-chicken niches are mostly generalists. This finding is in line with Dearlove *et al.* (2016) who suggested that generalists are more infectious to humans due to rapid host switching.

Our data shows chicken specialists to be the only niche restricted strains highly infectious to humans. Interpreting differential transmission based on host switching capabilities crucially depends on whether my chicken generalist indices are artefactual. If biological underpinnings cause chicken specialist lineages to be infectious, this could explain the higher transmission rates from poultry. Commonalities in immune response, gastrointestinal tracts, or the different handling of poultry compared to other meats could explain the higher infectivity. However, the high number of chicken specialists could also be due to the oversampling of chickens (4,147 samples compared to 716 cattle and 584 sheep). The high number of chicken restricted lineages is especially

striking as the other farm animals show a low number of specialists before and after transmission. However, the residual farm animals being sheep and cattle, which likely have similar gastro-intestinal features, would make the evolution of niche specialism less likely. As Sheppard *et al.* (2011b) speculate, the ecological factors of a farm environment are conducive to host switching. An increased potential for horizontal gene transfer through proximity enables closely related strains to colonise phylogenetically distant host animals. My findings indicate that chicken niche specialists must have a unique ability to colonise humans compared to other farm niche specialists. Beyond the investigation of the transmission bottleneck through strain composition, which would also be possible using older methodology, my source attribution allows me to attribute individual samples, which opens new lines of investigation.

2.3.6 The Fine-Grained Structure of Source Attribution can be Identified with Machine Learning

Past predictive efforts in source attribution based the assigned label on the genotype frequencies in the host reservoirs observed in sampling. However, sampled source distributions are not automatically accurate depictions of relative contribution to human infection. Lineages outcompeted by others in one niche could have increased fitness given a different environment, suggesting that rarely observed source lineages could disproportionally contribute to human infection. This dynamic has been exemplified in the chicken processing chain, where genes promoting an increased survival ability outside the gut, has led to an up-sampling of strains exhibiting the fitness promoting allele (Yahara, Méric, Taylor, Vries, et al. 2017). Similar evidence exists for the transmission bottleneck, where genes associated with human niche tropism leads to shifts in the composition of *C. jejuni* strains assemblages (Méric *et al.* 2018). With my

algorithm, which is capable of attributing sources to individual samples, the relative frequency changes in transmission can potentially be detected in a finer-grained source attribution.

The application of TreeBreaker (Ansari and Didelot 2016) allowed me to search for potential evidence of varying host affinities within the dataset. I focused on the phylogeny of CC-21 as a lineage found across multiple host reservoirs and commonly in human campylobacteriosis cases (Sheppard, Cheng, Méric, Haan, et al. 2014). The prediction of TreeBreaker shows a change in host association on a branch grouping of the cattle-associated ST-21 subgroup and the cattle-associated lineages ST-982 and ST-806 (Figure 2-6A). The observed host composition in this clade (asterisked in Figure 2-6A) is divergent from the residual samples of CC-21 consisting largely of chicken and sheep isolates. Furthermore, the clade under investigation showed an increased transmission to humans. Generally, CC-21 is over-represented in isolates from humans, potentially owing to its host generalist status. However, in the asterisked clade and the most related STs 982 and 806, I only observe expansion ranging between 1.7 and 3.6-fold, whereas the rest of CC-21 expanded 5.5 to 6.2-fold (Figure 2-6B).

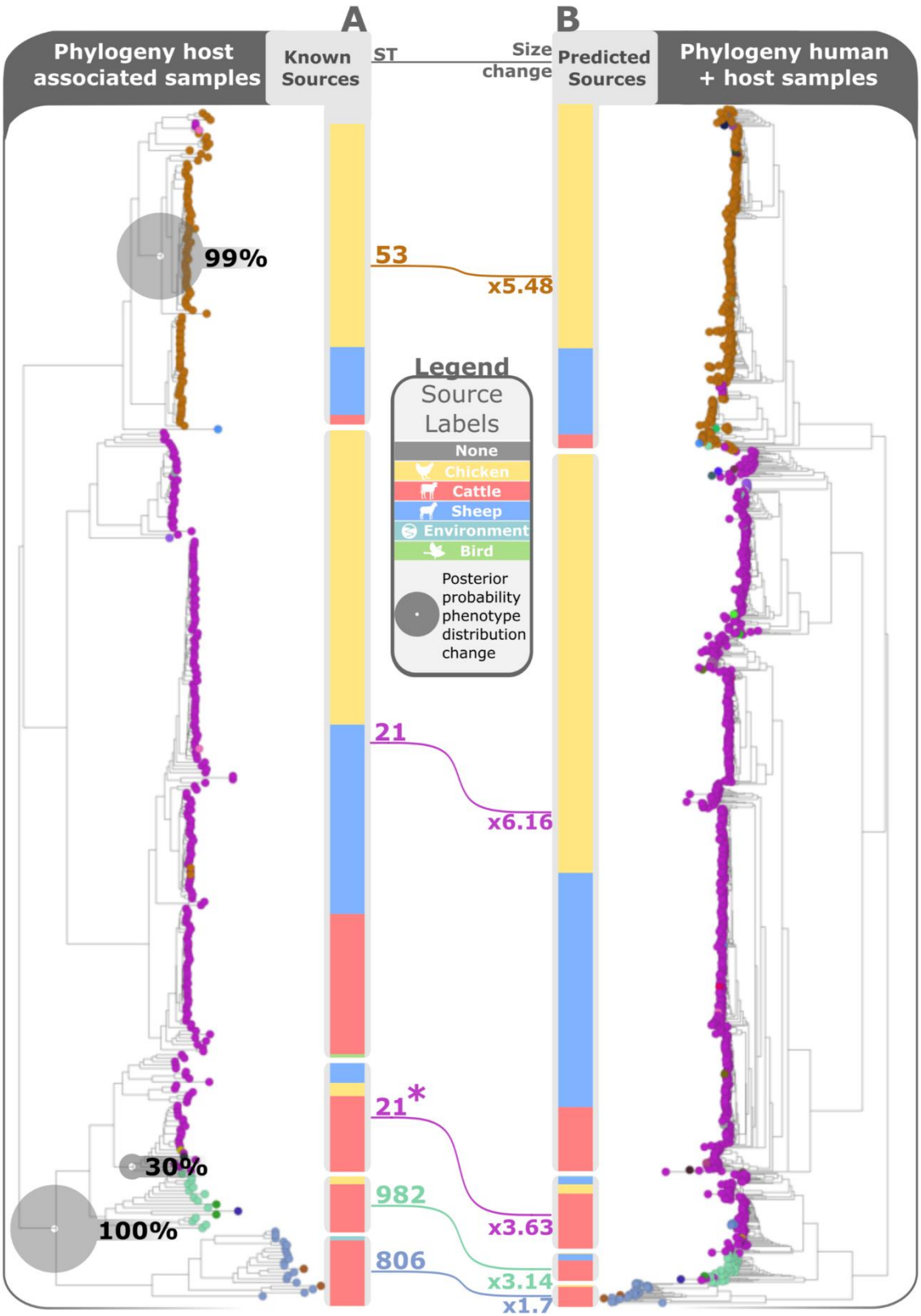


Figure 2-6 The varying host affinities of CC-21

Phylogeny of clonal complex 21 of host animal associated samples and (A) bar charts showing the known source distribution and human samples (B) alongside the predicted source distribution. The phylogeny is based on neighbour joining using hamming distance of the k-mers drawn from WGS. The connecting lines show the increase in frequency of the clades in human samples and the size of the grey circles show the posterior probability of a change in phenotypic distribution along the branches of the tree.

The apparent switch of host affinity within the lineage most abundant in human infections, also seemingly changed the ability of transmission to humans. A change in infectivity lead to a shift in predicted source composition of CC-21 in human infections compared to the same strain isolated from animals. Using previous methodology of assigning source, the discovery of this shift within CC-21 would have been impossible with potentially detrimental effects to understanding disease.

2.4 Outlook and Conclusions

With the ever-increasing wealth of available pathogen sequences grows the potential for broadening our incomplete understanding of zoonotic diseases given appropriate methodology. My analysis has revealed tree-based ensemble methods as fitting algorithms for classifying bacterial genetic sequences, which provides opportunity to improve the accuracy host source attribution for human Campylobacteriosis. The strength of this approach crucially hinges on utilising the full gradient of genomic differentiation offered by cgMLST and WGS data. Genetic differences specific to host reservoirs can be found in both core and accessory genes (Sheppard et al. 2013a) the use of which is subject to practical considerations. Having more computational power

at my disposal, could have enabled the use of all k-mers present across all genomes (here 109,675,176 unique kmers) whilst using multiple algorithms for cross-validation and bootstrap replication.

Beyond an increased ability for source attribution, by uncovering the fine-grained structure of genomic signatures of host specificity, aiSource offers tremendous opportunity to incorporate shifts in strain composition between samples taken from animal reservoirs and human isolates. This can facilitate the investigation of the ability of specific lineages to survive outside of the host sufficiently long to transmit to humans and the propensity to colonise human intestines when the opportunity arises (Yahara, Méric, Taylor, Vries, et al. 2017; Méric et al. 2018). Intuitively, this leads to questions about the genomic underpinnings of bacterial host adaptation, specifically the extent to which 'associated' genetic elements represent adaptations and whether the same genes and alleles enable colonisation of different host niches.

Improving on my approach, more abundant sampling, and subsequent incremental training of aiSource could provide additional improvement The algorithm's low computational requirements combined with a high prediction speed make it an excellent tool for prediction of large genomic datasets. Moreover, by using phylogeny-aware train/test splitting for assessing accuracy, the high predictive performance should prevail even when given new genetic variants. As implemented in the algorithm available on https://github.com/narning1992/aiSource, the classifier can readily be retrained given new data or different phenotype labels. AiSource thus has considerable potential for the deployment in an automated and continuous disease surveillance systems to reduce the burden of Campylobacteriosis that remains one of the most common food-borne illness in the world.

Chapter III

Genome wide association studies in

Campylobacter jejuni

Overview

The pathogen *Campylobacter jejuni*, the most common aetiological agent of foodborne gastroenteritis, is also a pervasive gut symbiont of farm animals. Its ability to switch hosts allows for increased transmission events and is common across multiple *C. jejuni* strains. Understanding the genetic underpinnings of host generalism is important for combatting Campylobacteriosis because the most common strains in human infection are also universally present in farm animals. I conducted a genome wide association study (GWAS) based on sequences from multiple studies found in PubMLST, utilising phenotypes based on the prediction of new source attribution methods, through which I aim to provide a more comprehensive analysis of *C. jejuni* niche adaption than previously possible. I identified a common fluoroquinolone resistance mutation which pre-adapts *C. jejuni* chicken samples for transmission to humans on a population level. The effects of the mutation previously uncovered in a moth model highlights the dangers of pervasive antibiotic use in food production beyond the rise of antimicrobial resistance. In both ruminant and chicken niches I uncovered genes that increase survival in the varied stressors of the food production chain, and invasion of epithelial cells increasing affinity to human hosts. I also identified polyphosphate pathway genes that are differentially associated between the ruminant and chicken niche. My results demonstrate the utility of large-scale meta-approaches to investigate pathogen evolution based on public databases. By establishing a robust methodology for combining heterogeneous studies, I hope to provide a first step towards streamlining the gain of evolutionary insight from the large wealth of publicly deposited genetic sequences, and to ultimately contribute to the development of automated GWAS.

3.1 Introduction

Campylobacter jejuni is primarily known as the pathogen responsible for the most common form of foodborne gastroenteritis worldwide (Kaakoush *et al.* 2015) but can also inhabit non-human hosts by acting as a symbiont in the gut of livestock animals and wild bird species (Burnham and Hendrixson 2018). The ability to colonise multiple hosts provides an advantage by increasing the habitat range, but may be limited by genetic constraints on the ability to adapt well to different host environments (Woolhouse, Taylor, and Haydon 2001), where the adaption to one host potentially comes at a fitness cost to another (Giraud *et al.* 2001). The trade-off between the competitive advantage conveyed by better adaption to one environment versus the benefit of switching niches is thought to create and maintain phenotypic diversity in a heterogeneous environment (Van Tienderen 1991). In *C. jejuni* the different evolutionary solutions manifest as a wide range of specialist and generalist strains (Sheppard, Dallas, MacRae, et al. 2009b; Sheppard, Dallas, Strachan, et al. 2009b; Sheppard, Colles, et al. 2010b; Sheppard, Cheng, Méric, de Haan, et al. 2014). As generalists are common contributors to human infection (Dearlove et al. 2016b), uncovering the genetic underpinnings of host switching is of paramount importance for alleviating the burden of Campylobacteriosis.

Supplied with a growing availability of genomic sequences, genome-wide association studies (GWAS) were conceived to examine human disease by uncovering causal variants as potential drug targets (Uffelmann *et al.* 2021). Since their inception GWAS have been widely used in bacteria to examine the genetic basis of a range of important phenotypes relating to severity of infection, antimicrobial resistance (i.e. Young *et al.* 2021; The CRyPTIC Consortium 2021; Earle *et al.* 2021) and host affinity (2013).

Sheppard et al (2013) were the first to apply GWAS to investigate host adaption of *C. jejuni* to cattle and chicken using 192 isolates of the generalist clonal complexes (CC) 21 and 45. The presence of the three *panBCD* genes responsible for Vitamin B_5 biosynthesis were found to be essential for survival in cattle, presumably because Vitamin B_5 is scarce in grasses but abundant in chicken feed. The genetic basis of biofilm formation was analysed by Pascoe *et al.* (2015) using 102 assemblies from CCs 21 and 45. The GWAS revealed genes in motility, capsule production, adhesion, glycosylation and oxidative stress as important in biofilm formation. Yahara *et al.* (2017) used 600 genomes of CCs 21 and 45 from different stages of the poultry processing chain. The study uncovered genes involved in formate metabolism, aerobic survival, oxidative respiration, and nucleotide salvage to enable *C. jejuni*'s survival from farm to fork. Buchanan *et al.* (2017) undertook the first GWAS investigation into pathogenicity, using 166 representative genomes to identify 25 genes as diagnostic markers. Vitamin B_5 biosynthesis pathway members were again found alongside genes involved in iron acquisition, and β-lactam antibiotic resistance. In the most recent GWAS, Epping and colleagues (2021), looked at host specificity on the basis of 490 genomes of animal, environmental and human origin identifying markers involved in genome maintenance and metabolic pathways. The five published GWAS offer a multi-faceted view on the mechanisms of niche adaption of *C. jejuni* using a study design tailored to the focal phenotype.

The utility of a "prospective" study design with carefully selecting, sequencing, and assembling isolates for analysis is not the only way of garnering insight into the phenotype under investigation. The deposition of sequences generated in past GWAS and similar studies onto public databases has amassed a wealth of publicly available *C.*

jejuni genomes, laying the foundation for new "retrospective" study designs. Looking at previous studies reveals that the potential of the 53, 097 *C. jejuni* assemblies listed on PubMLST (Jolley, Bray, and Maiden 2018) remains mostly untapped. The most extensive GWAS to date uses 600 samples (Yahara, Méric, Taylor, de Vries, et al. 2017) and three out of five studies focus on just two *C. jejuni* clonal complexes out of 42 listed on PubMLST. I aim to build upon past *C. jejuni* studies and combine the published data into a more comprehensive analysis of adaption to varying host environments. The retrospective approach offers the benefit of increased power due to bigger sample size with the ability to study multiple phenotype splits using the same methodology. Combining heterogeneous studies however comes with caveats, as differences in sequencing setup and assembly techniques can confound the analysis (Tom *et al.* 2017). I therefore aim to establish a robust protocol for bacterial GWAS based on public databases which can harvest the potential of the ever-increasing availability of genomes.

The large public databases at my disposal not only facilitate retrospective analyses with bigger sample sizes, but also allows me to pose different research questions using novel methodology. Attributing the source of *C. jejuni* infection was previously (Wilson et al. 2008b) based on 7-gene MLST (Maiden et al. 1998b). In MLST, the smallest predicted entity are entire STs. The development of a new source attribution method using cgMLST, comprising 1,343 genes, enabled source prediction at the level of individual samples (Arning *et al.* 2021 and chapter 2). I am therefore equipped with novel data that allows us to conduct original analysis, namely investigating the separate transmission chains from ruminant and chicken to humans. I additionally revisit the

question of niche adaption to ruminants and chicken in a bigger sample size. Comparing samples collected from ruminants and chicken uncovered genes in nucleotide salvage and chemotaxis regulators towards iron and phosphate as crucial for host specificity. I also report that mutations conferring resistance to antimicrobials also increase transmission from chickens to humans. The study of different evolutionary trajectories from two main sources of infection could allow for more bespoke public health interventions to alleviate the disease burden of Campylobacteriosis.

3.2 Methods

3.2.1 Dataset Collection and Preparation

The study presented here was based on the dataset used in Arning *et al.* (2021 and chapter 2) which can be accessed under_pubmlst.org/bigsdb?db=pubmlst_*Campy-lobacter*_isolates&page=query&project_list=102&submit=1. I strictly filtered all 20,314 available genomes to avoid confounding the analysis by the heterogeneous nature of the contributing studies (Tom *et al.* 2017). To prevent artefacts resulting from differing sequencing workflows I selected 10,111 genomes, by limiting all genomes to samples from bioproject accessions PRJEB2075, PRJEB4848 and PRJNA505131, which resulted in 7707 human, 1606 chicken, and 797 ruminant samples. I filtered the database down to these three studies to avoid batch effects due to sequencing chemistry or technology. The three studies have similar sequencing setups with read lengths ranging between 190 and 210 base-pairs and using Illumina HiSeq 2500 or 2000 as sequencer and containing only samples from the UK. Bioproject PRJNA505131 had no ruminant associated samples and was therefore removed for all phenotype splits with ruminant source samples.

3.2.2 Bioinformatics Processing

As differing assembly strategies potentially generate erroneous hits, all genomes were reassembled from scratch. The read data was obtained from the sequence read archive (Leinonen, Sugawara, and Shumway 2011) and adapter sequences trimmed using fastp (S. Chen et al. 2018). Reads were corrected with BayesHammer (Nikolenko, Korobeynikov, and Alekseyev 2013) before assembly with SPAdes (Bankevich *et al.* 2012) using the –careful flag. The resulting assemblies were aligned to 10 chromosome level assemblies of *C. jejuni* (accession numbers: ASM143234, ASM231296, ASM993964, ASM993966, ASM993970, ASM993972, ASM1320568, ASM1336377, ASM1336381, ASM1339499) using Ragout (Kolmogorov *et al.* 2014) and assembly errors were corrected using POLCA (Zimin and Salzberg 2020) both in default mode.

AiSource (Chapter 2, https://github.com/narning1992/aiSource, Arning *et al.* 2021) was retrained based on the core-genome multi-locus sequence typing (cgMLST) obtained from the source associated genomes using PubMLST (Jolley and Maiden 2010) with the sheep and cattle labels combined into a ruminant class, and pig, bird and environment classes combined into an “other” class (Cody et al. 2017b). The chicken, ruminant and other classes were used for training, and the human samples predicted as the “other” class were removed from the analysis as they were not of interest for the GWAS. The machine learner predicted 6374 chicken, 1480 ruminant and 290 “other” samples from the 8144 samples isolated from humans. Genomes were subsequently filtered as described below before use in the GWAS.

With contamination being a source of spurious hits, all contigs not originating from *C. jejuni* as predicted by Kraken (Wood and Salzberg 2014) were removed from the

assemblies. The statswrapper script from the BBMap suit (Bushnell 2014) allowed the removal of outliers in assembly characteristics, such as number of bases, number of contigs, N50, gap percentage, GC-content, read coverage. The read coverage was generated by mapping all reads to the reference genome NCTC11168 with Bowtie 2 (Langmead and Salzberg 2012) using the -k 1 and -very-careful flag. The depth tool within the Samtools package (H. Li *et al.* 2009) was used to generate a coverage from the reads. For every assembly characteristic listed above, genomes deviating more than 1.5 times the inter quartile range from either the first or third quartile were discarded. After generating 31-mers from all genomes using DSK (Rizk, Lavenier, and Chikhi 2013b) for use in the GWAS, all genomes diverging more than 1.5 times the inter-quartile range from the first and third quartile of the number of unique k-mers were removed. After quality filtering the dataset consisted of 7,211 genomes.

3.2.3 Study Design

To investigate the adaption of *C. jejuni* into different host niches, appropriate phenotype splits for the GWAS were chosen from the dataset.

- **Ruminant source vs ruminant human:** *C. jejuni* isolated from colonised ruminants (controls; n = 394) versus *C. jejuni* isolated from human infections predicted to originate from ruminants (cases; n = 553). I aimed to identify genetic variations allowing *C. jejuni* to transmit from their commensal lifestyle in sheep and cattle into human Campylobacteriosis.
- **Chicken source vs chicken human:** as above but focusing on isolates from colonised chicken (controls; n = 360) versus human infections predicted to be of chicken origin (cases; n= 2095).

- **Chicken source vs ruminant source:** isolates from colonised chicken (cases; n = 279) versus isolates from colonised ruminants (controls, n= 218). This revisits the question of genetic variants associated with different commensal environments.

3.2.4 GWAS

The GWAS was performed using the R package bugwas (https://github.com/sgearle/bugwas, Earle *et al.* 2016) that controls for population structure by using linear mixed models (LMM) with GEMMA (X. Zhou and Stephens 2012). To capture the genetic variability in the dataset, a presence/absence table of all unique 31-mer patterns was produced using k-mers generated from the assemblies by DSK (Rizk, Lavenier, and Chikhi 2013b). From the k-mer presence/absence table, a centred relationship matrix was calculated in Java which was (1) input into GEMMA and (2) to construct a phylogenetic tree via neighbour joining in the R package ape (Paradis, Claude, and Strimmer 2004). For additional, statistical, control of batch effects, the following variables were included as covariates: coverage of the genomes in reads, number of contigs, the contig N50, the scaffold L50, the average GC-content and the gap percentage of the genome. The bioproject accessions were included as factors, encoded via dummy variables. I used a 2% minor allele frequency threshold for GEMMA. The multiple testing corrected significance threshold was computed using the Bonferroni method to control the strong-sense family-wise error rate. A significance threshold of 0.05 was chosen which was divided by the number of tests defined by the unique 31-mer patterns observed in the phenotype split. This produced the following significance thresholds on a $-\log_{10}$ scale:

- Ruminant source vs ruminant human: $-\log_{10}(p) \geq 6.71$
- Chicken source vs chicken human: $-\log_{10}(p) \geq 7.17$
- Chicken source vs ruminant source: $-\log_{10}(p) \geq 6.77$

The phenotype split comparing ruminant and chicken samples showed inflated p-values in the quantile-quantile plots, so Treemmer (Menardo *et al.* 2018) was used to remove highly related samples from the dataset. The tree was trimmed at 97% of its original root-to-tip length by individually discarding leaves resulting in the case and control numbers shown under 3.2.3. The heritability of each phenotype split was computed by GEMMA using the LMM null model based on 31-mers.

3.2.5 GWAS interpretation

For interpretation of the significant k-mers, all unique 31-mers observed across all phenotype splits were iteratively mapped to multiple *C. jejuni* genomes. Using Bowtie2 with the -very-sensitive flag, all k-mers were first mapped to the *C. jejuni* NCTC11168 reference genome and all *C. jejuni* plasmid assemblies available on the National Center for Biotechnology Information (NCBI) assembly page. All k-mers not aligning with a mapping quality above 10 were then mapped via the same process to all 10 whole-chromosome level assemblies of *C. jejuni* available (accession numbers as listed in 2.2). The same process was repeated with a representative subset of all *C. jejuni* samples chosen with cd-hit (W. Li and Godzik 2006) using a cut-off of 97% identity (166 genomes). Finally, all remaining k-mers that did not map with a mapping quality of at least 10 were mapped against all 414 available complete *C. jejuni* genomes with genome annotation on NCBI assembly. If k-mers were still not mapped with a quality of 10 after 4 rounds of bowtie, BLAST (Altschul *et al.* 1990) was used with a query coverage of at

least 90 and an e-value of at least 0.1 and at least 95% identity against the same 4 datasets as described in the bowtie mapping. All hits that did not map with a quality of at least 10 were considered unmappable for the purpose of this study.

The plots depicting the GWAS results, such as the Manhattan plots, quantile-quantile plots and forest plots were generated in Python 3.8 using the matplotlib (Hunter 2007) and seaborn libraries (Waskom 2021). The trees depicting k-mer presence/absence patterns were generated in Microreact (Argimón et al. 2016b). Individual genome-wide significant hits were plotted using the DNA Features viewer library (Zulkower and Rosser 2020) in Python with genome annotations downloaded from NCBI assembly. The k-mers were aligned to the reference genomes downloaded from NCBI using Mafft (Katoh *et al.* 2002) with the -auto function. K-mers were merged when they overlapped by more than 15 bases and their overlaps and their directionality of effect as measured by the regression coefficient beta were identical. To confirm the hits are not the result of systematic differences between the three different bioproject study designs, such as uncontrolled batch artefacts, I repeated the GWAS by performing three pairwise comparisons between all bioproject accessions, treating the bioproject as the phenotype. The effect size and direction of the LMM were recorded as z-scores of the beta value (the beta estimate divided by the standard error).

For the comparison of resistance conferring mutations in the NCTC11168 genome, I used AMRFinderPlus (Feldgarden *et al.* 2021) with default command line options.

3.3 Results

Starting with 20,314 *C. jejuni* genomes publicly available from PubMLST, I filtered the dataset to 10,148 samples originating from three studies with similar sequencing

characteristics to avoid confounding batch effects. I further filtered outliers based on sequence characteristics, resulting in 7,211 genomes, all phenotypes considered.

3.3.1 Isolates from ruminant versus human isolates predicted to come from ruminants

We first compared isolates drawn from ruminants and human *C. jejuni* samples predicted to be transmitted from contaminated ruminant meat or milk. The sample heritability of the phenotype was 25.4 % +/- 7.5 (one standard error). As seen in Figure 3-1 no k-mer passes the genome-wide significance threshold at a -$\log_{10}$ p-value of 6.7, potentially due to the fewer samples available in this split (n = 947) compared to other splits under investigation.

Figure 3-1 Ruminant source vs ruminant human hit 1.

A detailed view of the most significant hits of my GWAS study comparing source associated ruminant samples with human samples predicted to come from ruminants. A) The Manhattan plot of the GWAS. B) A zoom in around the most significant hit shows the genome annotation of the NCTC11168 in closer proximity to the hit. C) A forest plot shows the beta of the association across different bioproject accessions. D) An alignment of all significant k-mers with the NCTC11168 reference genome with the z-scores of beta listed on the left.

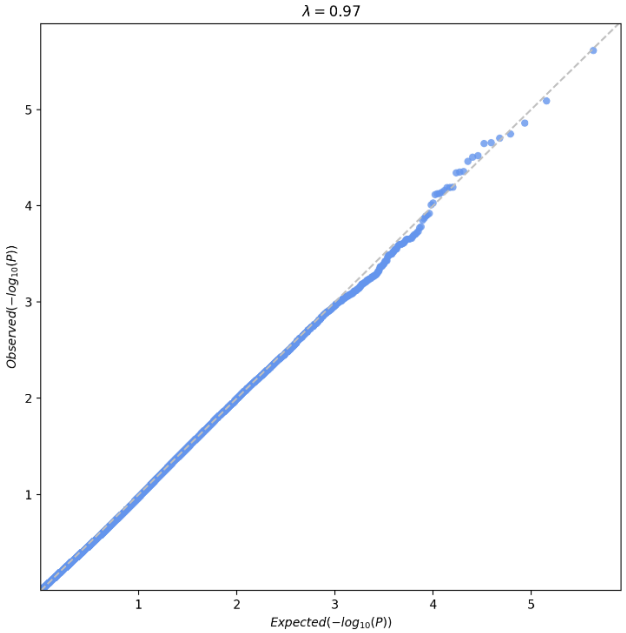


Figure 3-2 Qq-plot Ruminant source vs ruminant human.

The quantile-quantile plot of the genome-wide association study comparing source associated ruminant samples with human samples predicted to come from ruminants. The -log10 p-values of all k-mer patterns analysed in the association study are compared with quantiles of a uniform probability distribution to an inflation of significance.

Figure 3-1 shows the region of the most significant 31-mers, which had a -$\log_{10}$ p-value of 5.6 and a z-score of the beta of -4.7, with Figure 3-2 showing the associated qq-plot. The negative directionality of effect signifies an association with samples drawn from ruminants. The k-mers are mostly identical to the intergenic region on NCTC11168 at positions 878453 to 878500 with a few single nucleotide mutations in individual sequences. The most significant k-mer with a positive directionality of effect (z-score of beta = 1.83) indicating association with human samples, introduced an 8bp gap into the alignment.

3.3.2 Isolates from chicken versus human isolates predicted to come from chicken

C. jejuni samples isolated from chicken were compared with human-isolated samples predicted to be transmitted via contaminated chicken. This test of association between chicken samples and human *C. jejuni* isolates transmitted via chicken revealed that 22.78 % +/- 3.4 of the phenotypic variability could be attributed to the genotype. Four distinct k-mer patterns pass the significance threshold of -$\log_{10}$ p-value of 7.17. Only the most significant hit at a -$\log_{10}$ p-value of 10.7 could be mapped above the quality threshold of 10 (Figure 3-3). The qq-plot of the phenotype shown in Figure 3-4.

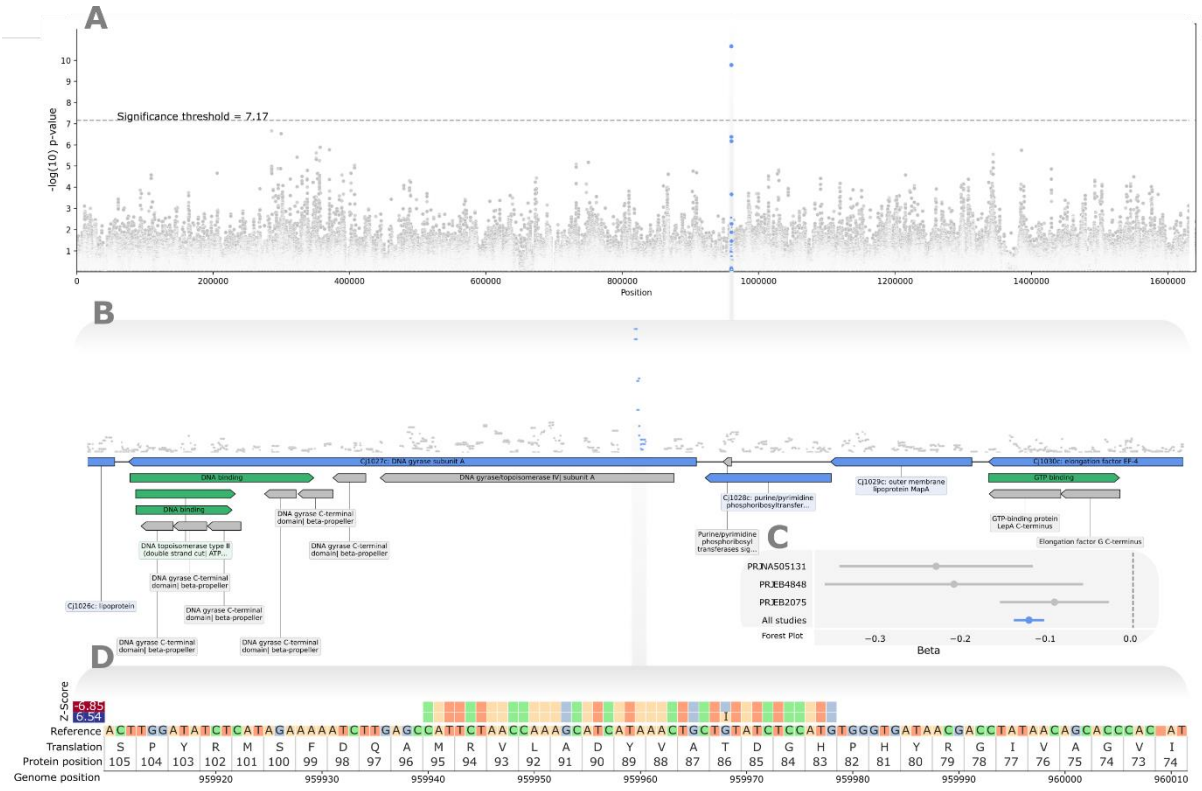


Figure 3-3 : Chicken source vs chicken human hit 1.

A detailed view of the most significant hits of my GWAS study comparing source associated chicken samples with human samples predicted to come from chicken. For a more detailed explanation of the individual panels see the caption of Figure 3-1. Additionally, panel D) here shows an amino acid translation of the protein the k-mer maps to.

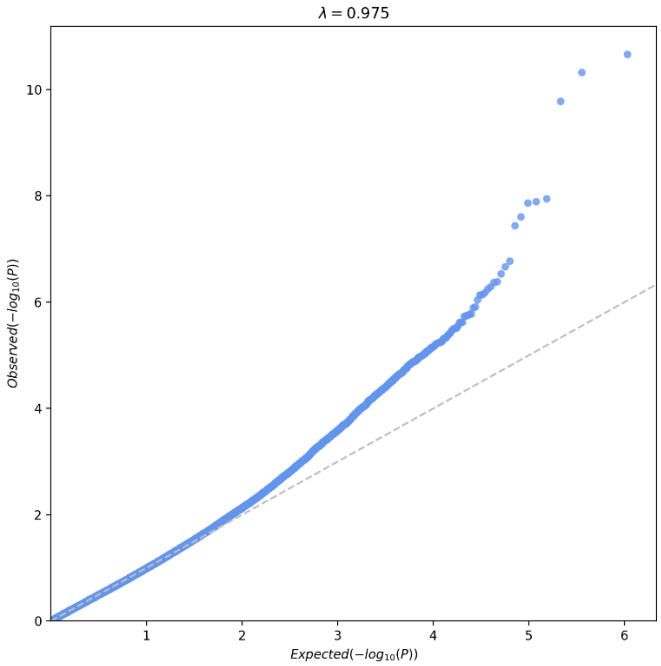


Figure 3-4 Qq-plot of Chicken source vs chicken human

The quantile-quantile plot of the genome-wide association study comparing source associated chicken samples with human samples predicted to come from chicken.

3.3.3 The k-mer associated with the most significant pattern mapped within the gene named DNA gyrase subunit A (*gyrA*), also known as DNA topoisomerase 2. The reference version of the gene is most significant and associated with chicken isolates with a z-score of beta of -6.85, whereas a G to A substitution at gene position 247 is linked to human isolates with a z-score of beta at 6.54. Isolates from ruminants versus isolates from chicken

Comparing chicken to ruminant isolates revealed a higher sample heritability than all other splits at 60.3% +/- 6.1 %. Two patterns passed the corrected threshold of a -$\log_{10}$ p-value of 6.7, with the top hit mapping best to assembly CP013116.1 which was generated by sequencing *C. jejuni* strain T1-21.

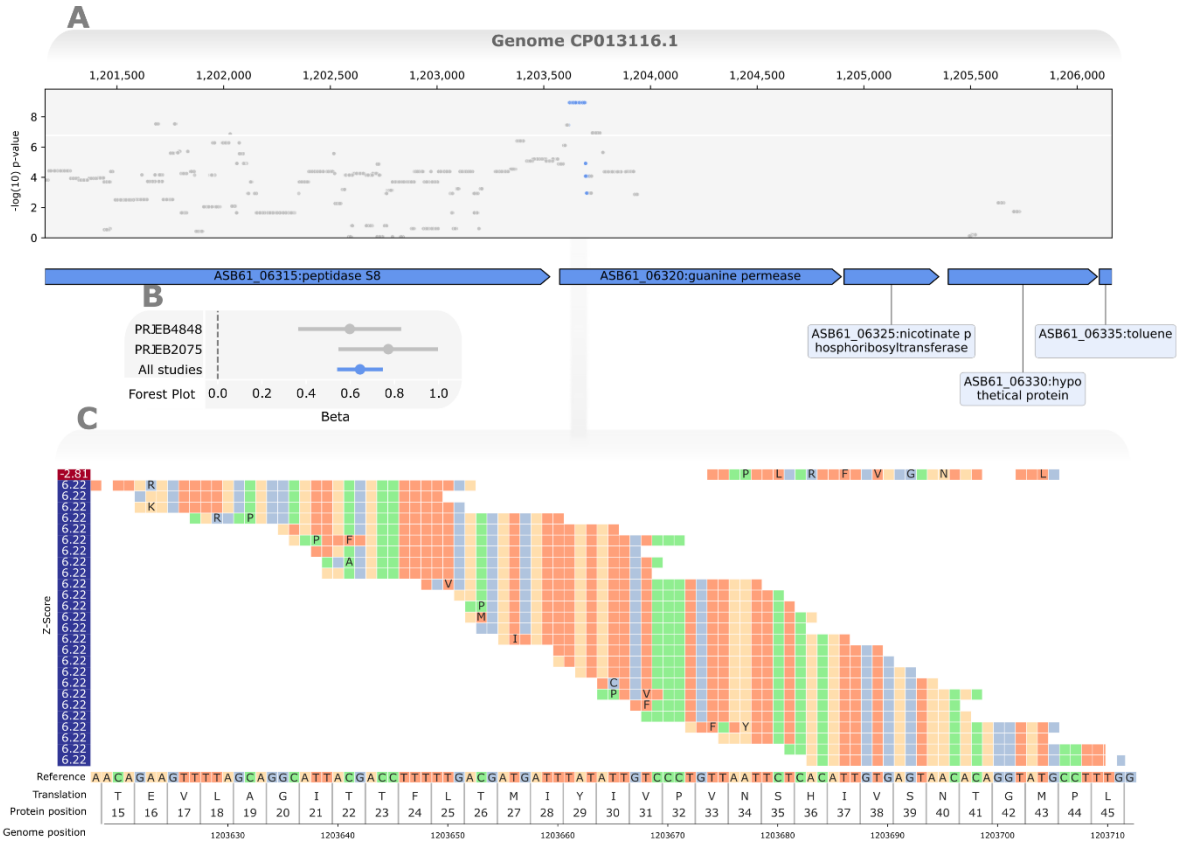


Figure 3-5 Ruminant source vs chicken source hit 1

A detailed view of the most significant hits of my GWAS comparing source associated chicken samples with source associated ruminant samples. The Manhattan plot for all hits on genome CP013116.1 is omitted, as only few k-mers map onto this genome. For a more detailed explanation of the individual panels see the caption of Figure 3-1.

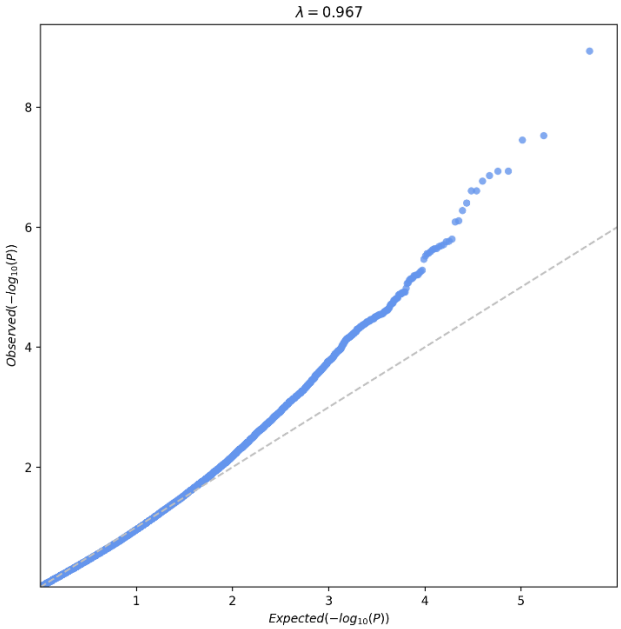


Figure 3-6 Qq-plot of ruminant source vs chicken source

The quantile-quantile plot of the genome-wide association study comparing source associated chicken samples with source associated ruminant samples.

The pattern shown in Figure 3-5 has a -$\log_{10}$ p-value of 8.9 and a beta z-score of 6.22 which suggests an association with ruminants. The qq-plot of the GWAS is shown in Figure 3-6. The hit maps to the non-reference *C. jejuni* genome CP013116.1 within the gene producing the protein ASB61_06320.

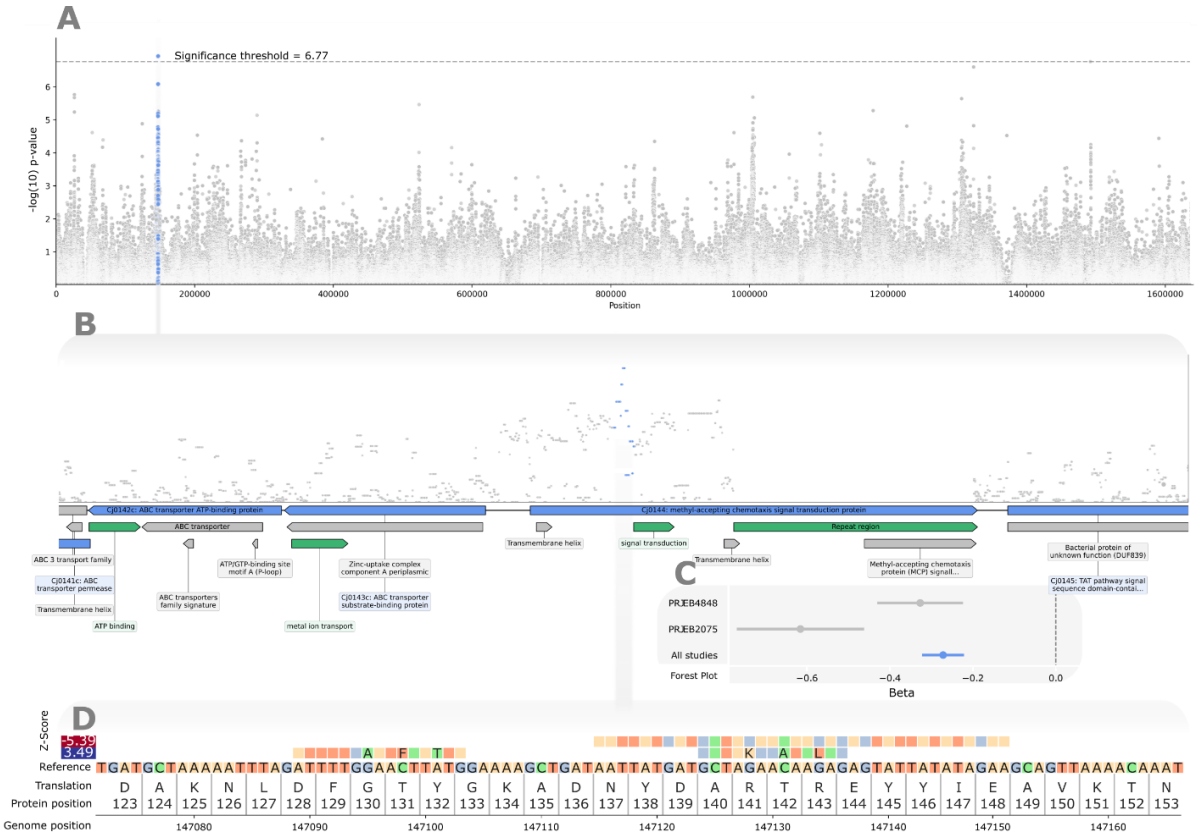


Figure 3-7 Ruminant source vs chicken source hit 2

A detailed view of the second most significant hits of my GWAS study comparing source associated chicken samples with source associated ruminant samples. For a more detailed explanation of the individual panels see the caption of Figure 3-1.

The second most significant hit at a -log10 p-value of 6.9 has a z-score of beta of -5.46, as seen in Figure 3-7, which indicates an association with chickens, as opposed to ruminants. The associated k-mer maps on the reference genome NCTC11168 within the gene cj0144.

3.4 Discussion

3.4.1 Glutamine Uptake Genes Associated with Transmission from Ruminants to Humans, but Did Not Reach Genome-Wide Significance

The most significant k-mer in the comparison between isolates from ruminants and human isolates predicted to come from ruminants maps between the genes for Cj0939 and Cj0940. The proximity to both genes potentially indicatges regulatory function in one or both genes. In the following I adapt non-capital italic names for genes and capital non-italic names for the corresponding proteins. Cj0939 is a hypothetical protein with no known function, making it difficult to draw conclusions about its role in pathogenicity. Cj0940 is known as the glutamine ATP-binding cassette (ABC) transporter permease protein (GlnP). Bacteria use ABC transporters to move nutrients across the cellular membrane to scavenge substrates from within the host (Tanaka *et al.* 2018). *C. jejuni*'s growth is reliant on the uptake of amino and keto acids from the host microenvironment, due to the inability to use sugars as a carbon source (Velayudhan and Kelly 2002). Glutamine is used by *C. jejuni* to generate glutamate via hydrolysis which in turn can be decomposed into carbon (Hofreuter, Novik, and Galán 2008). Glutamine is also vital within the host as a nitrogen source through the deamination of glutamate (Lin *et al.* 2009). The ability of *C. jejuni* to scavenge glutamine from the host gut by this mechanism was shown to be crucial for pathogenesis of *C. jejuni* (Pei *et al.* 1998; Leon-Kempis *et al.* 2006, Lin *et al.* 2009).

Evidence for the involvement of GlnP in *C. jejuni*'s ability to survive the hostile gut environment is comes from functional analysis of *glnP* itself and through the study of glutamine scavenging within humans more broadly. *GlnP* was upregulated 2-fold or more in *C. jejuni* under hyperosmotic stress (Cameron *et al.* 2012), which can occur

during passage of the intestine or food processing (Cameron *et al.* 2012). De Vries *et al.* (2017) additionally show that mutants of *glnP* have reduced colonisation ability in chicken. The involvement of *glnP* in transmission may be twofold, by enabling survival of *C. jejuni* on the ruminant carcass and through protection of the bacterium in the gut. Hofreuter, Novik and Galan (2008) show that growth of the highly pathogenic strain 81-176 in the presence of glutamine and asparagine is possible which was an unviable environment for the reference strain NCTC11168. Glutamine transport is likely a redundant mechanism as PaqP was experimentally identified as another ABC transporter permease *in C. jejuni* (Lin *et al.* 2009). Paqp appears to be upregulated in bacteria during human cell infection (Gaynor *et al.* 2005) and its deletion resulted in increased recovery of bacteria of epithelial cells and resistance to aerobic and organic peroxide stresses. With the proximity of the most significant hit to *glnP* and its involvement in human infection, it is conceivable that regulation occurs within the discovered region. Although the association does not pass the significance threshold, being the most significant hit in the analysis supported by the experimental evidence is suggestive of an involvement in virulence. My analysis offers some support for further experimental validation for the ability of the region 878,453 to 878,500 to influence transmission to humans, possibly through the regulation of *glnP*

3.4.2 Fluoroquinolone Resistance Mutations are Associated with Transmission from Chickens to Humans

The most significant hit in the comparison of isolates from chicken and human isolates predicted to be from chicken is a point mutation within the *gyrA* gene. The point mutation causes a change from threonine to isoleucine at amino acid position 86 and is the most common mechanism of fluoroquinolone (FQ) resistance in *C. jejuni* (Piddock

2003; Sproston, Wimalarathna, and Sheppard 2018). DNA gyrase precedes the DNA polymerase during DNA replication by negatively super-coiling the tightly wound DNA (Reece and Maxwell 1991). FQ disrupts replication by binding the DNA onto the gyrase resulting in double strand breaks of the genome eventually causing the death of the bacterium (Willmott *et al.* 1994). The *gyrA* mutation is believed to convey resistance to FQ by lowering the binding affinity of FQ to the DNA-gyrase A complex (Aldred, Kerns, and Osheroff 2014). The involvement of *gyrA* in FQ-resistance is well established, however the mechanism of involvement in pathogenesis has only recently been revealed.

At first glance, the association between infection and antimicrobial resistance may appear to be a sampling artefact. As antimicrobial resistance encumbers routine treatment of gastroenteritis, protection from antibiotics could be a factor contributing to hospitalisation and thereby increased likelihood of *C. jejuni* sampling. Past studies however suggest that a side effect of the resistance-causing mutation in *gyrA* is to increase the fitness of *C. jejuni* within the host, which would also explain why it is the only antimicrobial resistance conferring mutation discovered here. Luo *et al.* (2005) first demonstrated that colonies carrying the *gyrA* mutation outcompeted FQ-sensitive strains when both were injected into chicken in the absence of antimicrobials. Whelan and colleagues (2019) extended this finding to *Galleria mellonella* moth larvae, demonstrating an increased invasion of epithelial cells and increased lethality in strains carrying the *gyrA* mutation. Whelan *et al.* propose that *C. jejuni* uses DNA supercoiling for controlling transcription (Shortt *et al.* 2016), since it lacks the intricate gene regulation mechanisms of other gastric pathogens (Jagannathan, Constantinidou, and

Penn 2001). The two-component regulatory system FlgRS is controlled by relaxation of DNA supercoiling and thus indirectly by *gyrA*. Relaxed supercoiling reduced FlgR expression and increased FlgS expression leading to decreased motility (Shortt *et al.* 2016). Both flagella genes are known to be important factors in the ability of *C. jejuni* to colonise the host intestine (Hermans *et al.* 2011). Decreasing motility was shown to lead to enhanced biofilm formation under aerobic conditions (Whelan *et al.* 2019) and relaxed supercoiling has also been demonstrated to increase invasion of human endothelial cells (Scanlan *et al.* 2017). The *G. mellonella* model showed that relaxed supercoiling can be attributed to the resistance causing *gyrA* mutation also resulting in an associated increase in invasion of epithelial cells and virulence (Whelan *et al.* 2019). The experimental validation of mechanism of the *gyrA* mutation at position 247 in transmission in the moth model, corroborates the association discovered here. FQ-resistance conferring mutations can influence transmission in two ways: By protecting *C. jejuni* from the oxidative stress during poultry processing by biofilm formation and increasing virulence through enhanced entry into epithelial cells (Whelan *et al.* 2019). Enhanced biofilm formation could also be the mechanism of resistance to FQ, which was shown to be the case for urinary tract infection causing *Escherichia coli* (Oliveira, Dias, and Pomba 2014). Biofilms composed of environmental DNA, as found in the biofilm of *gyrA* mutants (Whelan *et al.* 2019), safeguard the bacteria from contact with antibiotics. The *gyrA* gene could thereby reinforce the resistance conferred by decreased binding affinity to FQ, or conceivably be the main mechanism of resistance to FQ. The GWAS carried out here, appears to extend the conclusions drawn from moth models by Whelan *et al.* (2019) to the population level: The FQ resistance-causing *gyrA* mutation

pre-adapts *C. jejuni* for transmission to humans by increasing survival in food processing and entry into epithelial gut cells.

The use of FQ in poultry is already controversial, due to the rise in resistant strains, particularly as FQ is the most common treatment of foodborne gastroenteritis (Sproston, Wimalarathna, and Sheppard 2018). However, the dual effect of the *gyrA* mutation may exacerbate the detrimental effect of pervasive antibiotic use during food production. The increased fitness of FQ resistant strains in chickens when co-inoculated with susceptible strains suggests the substitution pre-adapts *C. jejuni* to chicken. With the nucleotide change increasing survival on the chicken carcass and invasion of human endothelial cells as previously outlined, the single nucleotide change may increase fitness at multiple steps of the transmission chain. Whereas FQ is still widely used in UK poultry production (European Food Safety Authority and European Centre for Disease Prevention and Control 2020), the US banned FQ in poultry production in 2005 (L. B. Price et al. 2007). As the use of FQs for sheep and cattle continued to be allowed, the US experienced a rise in prevalence of FQ resistant strains in ruminants (Tang *et al.* 2017; Xia *et al.* 2019). If the *gyrA* mutation is indeed responsible for both antimicrobial resistance and increased transmission to humans, there should be more relative transmission from ruminants than chicken in the US compared to the UK. It is interesting to note that, when I used aiSource (Arning *et al.* 2021 and chapter 2) to compare the predicted source of all available UK (n=15,050) and US cgMLST samples (n=437), I observe a higher proportion of US samples attributed to ruminants (49.9%) compared to the UK (28.2 %). A better comparison would be the pre-2005 source attribution compared to samples from after the FQ ban, however this was not possible due to the

scarcity of sequences before 2005. The evidence that the FQ resistance-conferring mutation increases *Campylobacter* transmission to humans adds weight to the argument for banning FQ, not only for chicken production, but across all farm animals.

3.4.3 GWAS Comparing Chicken and Ruminant Isolates Reveal Nucleotide Salvage and Chemotaxis Towards Iron and Phosphate as Host Associated Factors

The most significant hits in the comparison of isolates from chicken and ruminant map to the gene producing ASB61_06320 on the genome CP013116.1. The biological interpretation of this hit is difficult as the annotation of genome CP013116.1 is less complete than *C. jejuni* reference genome annotation. Additionally, experimental validation is lacking to illuminate the mechanism by which the mutation can increase fitness in one niche is lacking. However, there is suggestive evidence for the involvement of the associated hits in guanine uptake which could interact with the polyphosphate pathway as is laid out below.

The protein produced by the gene the most significant hit maps to, ASB61_06320, is a guanine permease and the mapped k-mers show multiple mutations associated with amino acid changes. When using BLAST (Altschul *et al.* 1990) against the better annotated *C. jejuni* proteome, ASB61_06320 shows closest similarly to the protein encoded by *cj1369* (identity = 90%, e-value = 0.044). The gene *cj1369* is part of the 3-gene nucleotide salvage operon found to be disease-associated in a GWAS study of *C. jejuni* (Yahara, Méric, Taylor, de Vries, et al. 2017). Yahara and colleagues stress the importance of nucleotide supply for replication, transcription, and translation albeit with the mechanism being unclear.

The analysis of purine biosynthesis genes *purF* and *apt* have shown purines to be important for intracellular epithelial survival by Cameron *et al.* (2015). The availability of purines was linked to the ability to survive osmotic and oxidative stress, due to their involvement in the polyphosphate pathway (Cameron *et al.et al.* 2015). Interaction with the polyphosphate pathway could explain the association of ASB61_06320 with host specificity as the protein was found to be differentially expressed in a knock down of the *C. jejuni* polyphosphate kinase (*ppk*) (Chandrashekhar *et al.* 2015). The *ppk* mutant was unable to survive osmotic shock and nutrient scarcity and was further incapable of colonising chicks, indicating a role in regulation of host affinity (Candon *et al.* 2007). The region discovered here might therefore be causal for a divergent ability to colonise ruminants and chicks through the interaction with polyphosphate, but the association invites further experimental validation.

The most second most significant hit in the comparison of isolates from chicken and ruminants maps to within the *cj1044* gene. Transcription of cj0144 produces a putative methyl-accepting chemotaxis signal transduction protein also known as transducer like protein 2 (Tlp2). In *C. jejuni*, transducer-like proteins are sensors of environmental stimuli through chemotaxis and energy taxis (Marchant, Wren, and Ketley 2002). The tlp2 knockout in C. jejuni led to decreased chemotaxis towards aspartate, pyruvate, inorganic phosphate, and iron and resulted in a decreased ability to colonise chicken intestines (Chandrashekhar et al. 2018). The gene was shown to be activated by inorganic phosphate and oxygen-reduced iron (Chandrashekhar et al. 2018). The oxygen-scarce gut microenvironment makes the reduced iron forms Fe2+ and Fe3+ the most readily available source of iron for gastric pathogens (Naikare et al. 2006). Iron

supply is vital for C. jejuni as a cofactor for enzymes and for the synthesis of iron sulphur proteins and cytochromes (Bencharit and Ward 2005). Buchanan et al. (2017) already found iron acquisition genes to be significantly associated with clinically related C. jejuni subtypes. Chemotaxis towards iron within the gut is vital as the host releases iron away from the bacteria (Naikare et al. 2006). Using ChIP assays to show interaction with iron uptake proteins in C. jejuni, Butcher et al. (2012) uncovered the alkaline phosphatase cj0145 producing the protein PhoX and discovered tlp2 to be activated by the same regulator in the presence of iron. PhoX uses iron as a cofactor to acquire inorganic phosphate through removal from phosphor-organic substrate (Yong et al. 2014). C. jejuni uses inorganic phosphate as the main source for the ppk mediated synthesis of polyphosphate. Tlp2 seems to mediate both iron acquisition by chemotaxis and, in tandem with PhoX, the generation of polyphosphate by first moving toward and then acquiring inorganic phosphate.

The importance of both iron and phosphate for C. jejuni is underlined by a study where iron and phosphate transport genes were highly expressed in chicks compared to in vitro cultures (Taveirne et al. 2013). Both hits discovered in my source associated GWAS indicate the importance of poly-phosphate generation for adaption to the chicken and ruminant niche. Polyphosphates are involved in hyperosmotic stress survival, carbon starvation and intracellular survival in epithelial cells (Candon et al. 2007). Phosphorus occurs in animal feed primarily in the form of phytates from which inorganic phosphate can be captured through phytase (Reddy, Sathe, and Salunkhe 1982). The enzyme is abundant in ruminants, but largely absent or inactive in the gut on monogastric animals such as chicken (Humer, Schwarz, and Schedle 2015). The difference in the availability

of phosphorus in chicken and ruminant intestines could explain why genes involved in chemotaxis towards phosphates and polyphosphate accumulation are the most significant associations in my analysis.

3.5 Conclusions and Outlook

Combining the sequences of multiple studies deposited on public databases has helped me shed light on the host adaption of *C. jejuni*. Genes improving survival through the processing chain and invasion of epithelial cells seem to be a common factor of transmission from ruminant and chicken sources to human. The same two factors also appear to be involved in pre-adapting chicken-derived *C. jejuni* for transmission to humans by the FQ resistance-causing *gyrA* mutation. As seen in Figure 3-8 only comparing chicken sources and human samples attributed to come from chicken shows a highly significant interaction between FQ resistance and phenotype. Other known antimicrobial resistance mutations also show no significant interaction with any phenotype split, indicating that FQ use in chicken in particular is problematic and should be replaced with other antimicrobials. Generally, my findings highlight the danger of pervasive antimicrobial use in animal husbandry beyond the alarming rise of antimicrobial resistance.

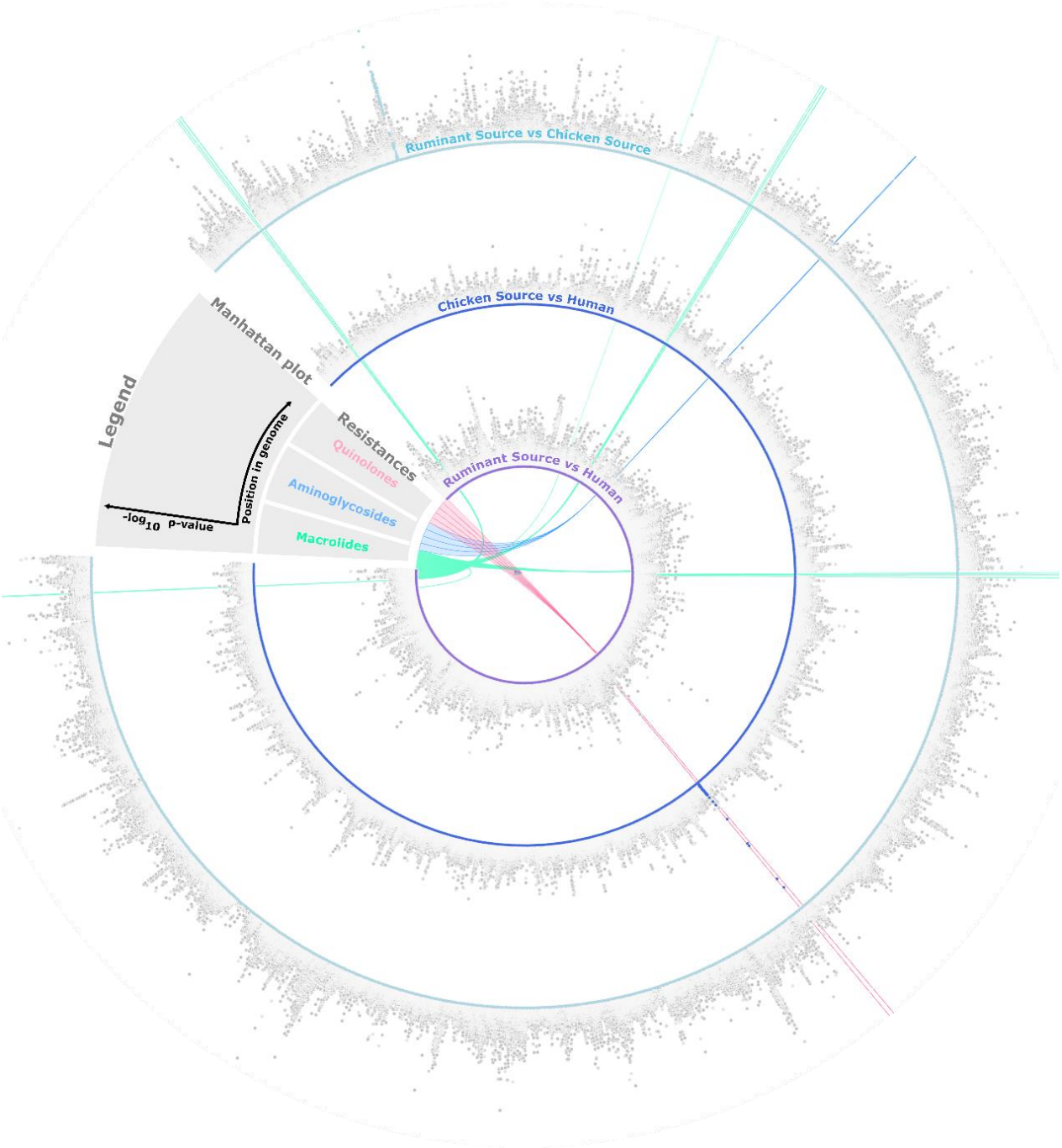


Figure 3-8 Antibiotic resistance across all GWAS studies

The inner circle shows known mutations conferring resistance to quinolones, aminoglycosides, and macrolides. The three outer rings show the Manhattan plots of the ruminant source vs ruminant human, chicken source vs chicken human and chicken source vs ruminant source GWAS successively. Significant associations are coloured in

the respective colours of the split, as also shown in the rings underneath the Manhattan plot.

In investigating host association within farm animals, I found genes involved in the polyphosphate pathway to be crucial for niche adaption due to a difference in phosphate availability within the chicken and ruminant gut. Our findings were only made possible through the generosity of researchers pursuing similar lines of investigation. It shows the great potential of large, well-curated and accessible public sequence databases for broadening the understanding of pathogen evolution. By using a careful approach of reassembly followed by an extensive regimen of filtering, and statistical control via inclusion of covariates, I provide a robust approach for combining heterogeneous sequencing collections for GWAS. As aiSource can be retrained using different labels for prediction, my approach can easily be repurposed for investigating other phenotype splits using cgMLST. I hope to provide a first step towards performing automated GWAS to investigate any trait or disease of interest using public sequencing databases. This work should facilitate access to the tremendous potential that largely lies dormant within the ever-increasing wealth of bacterial genomic data.

Chapter IV

Using Machine Learning and Bayesian Model Averaging to Analyse COVID-19 Risk

Overview

The emergence of the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) in 2019 prompted an urgent global demand for insight into the hitherto unknown coronavirus disease (COVID-19). The linkage of large scale comprehensive prospective cohort studies like UK Biobank (UKB) with COVID-19 test data provided an important tool to bridge the lengthy data collection step required for big data analysis. A COVID-19 risk assessment was particularly important to help overburdened health systems to target their limited medical attention and resources. This requirement was addressed by an abundance of studies utilising UKB data investigating varying hypotheses and presenting multiple assemblages of independent risk factors. Due to the intricacy of the causal chains involved in human disease, risk assessment crucially depends on variable choice, possibly explaining the divergent outcomes based on the same underlying database. Here I present an agnostic and comprehensive risk assessment through Bayesian model averaging that includes the full scale of available biomarkers in UKB. By also providing a machine learning based risk prediction, I can both provide analysis of risk and reassess previous efforts in using machine learning for inferential modelling. I find variables capturing socio-economic standing, ill health, and ethnicity to be important determinants of hospitalisation and severe forms of the disease. My susceptibility analysis emphasises the importance of housing arrangements influencing within-household transmission as the main driver of COVID-19 positivity. Using the full extent of information available it is possible to shed light on the plethora of published COVID-19 risk assessments based on UKB data and show that the discovery of risk factors through machine learning feature importance is questionable without further

analysis. As my algorithm can be readily reimplemented using other binary labels, I hope to provide a valuable tool for shedding light on any disease of interest using pre-existing large prospective cohort studies.

4.1 Introduction

The emergence of SARS-CoV-2 as an aetiological agent of the coronavirus disease that emerged in 2019 presented unparalleled challenges for healthcare systems globally owing to its high infectivity, virulence and asymptomatic transmission (Bi *et al.* 2020; Furukawa, Brooks, and Sobel 2020). There is an extensive clinical spectrum associated with coronavirus disease (COVID-19) infection including asymptomatic carriage, mild upper respiratory tract disease, and severe viral pneumonia (Huang et al. 2020; F. Zhou et al. 2020). The wider, pro-inflammatory, tissue invasiveness of SARS-CoV-2 has led to cardiovascular complications, renal injury, gastrointestinal disease and spread to the central nervous system with the potential of long-term persistence of effects (Bohn *et al.* 2020; Raveendran, Jayadevan, and Sashidharan 2021). After appearing in late 2019, SARS-CoV-2 became a global public health emergency that swiftly evolved into the greatest economic, political, and social crisis of the second decade of the 21st century (Rashedi *et al.* 2020). Numerous disciplines shifted their focus to combatting the global pandemic, which prompted an urgent demand for data to describe a previously unknown disease.

Understanding and predicting risk for severe forms of COVID-19 became an early goal to allocate resources and attention whilst the pandemic overburdened medical facilities (Sokolowska *et al.* 2020). Such a pre-hospitalisation risk assessment allowed the identification of potentially vulnerable patients, enabling the designation of groups for

shielding and early vaccine delivery (Ho, Celis-Morales, *et al.* 2020). With a variety of symptoms and the wide clinical spectrum observed, a wealth of heterogeneous data is required for understanding risk (Brotman *et al.* 2005). High quality data is arduous to collate under normal circumstances and especially difficult to obtain on timescales required to tackle an active pandemic. Fortunately for COVID-19 research, large publicly available prospective cohort studies like UK Biobank (UKB) (Sudlow *et al.* 2015) had been specifically designed to study disease and its associated risk. Dynamic linkage with COVID-19 test results and hospitalisation data by Armstrong *et al.* (Armstrong *et al.* 2020) facilitated COVID-19 research in UKB early in the pandemic. UKB is particularly well positioned for studying COVID-19 risk as participants age ranges from 50 to 87 and incidence of severity increases with age (Bi et al. 2020; Q. Li et al. 2020; Huang et al. 2020). Linkage by UKB to other COVID-19 data including mortality and hospital records, together with its wide array of epidemiological, medical, and genetic records, enabled numerous studies to identify risk factors of COVID-19 infection, hospitalisation, and mortality (Table 4-1).

Table 4-1 Table of COVID-19 studies based on UKB data

Study	Methods	Phenotypes	Associated Independent variables	Associated Comorbidities
Amin and Drenos 2021	GWAS summary statistics	Susceptibility Hospitalisation	No association found	
Anderson *et al.* 2021	Poisson regression	Hospitalisation Mortality		Venous thromboembolism
Atkins *et al.* 2020	Logistic regression	Hospitalisation Mortality		Fall or fragility fractures Coronary heart disease Type 2 diabetes Asthma Kidney disease Dementia COBD Pneumonia Depression Atrial fibrillation
Aung *et al.* 2020	Logistic regression and Cox models	Susceptibility	BMI Waist circumference Systolic blood pressure Serum glucose Serum glycated haemoglobin LDL cholesterol HDL cholesterol Triglycerides	
Batty, Deary, and Gale 2021	Cox regression	Mortality	Cognitive function	

G. D. Batty *et al.* 2020	Logistic regression	Hospitalisation	Education levels Deprivation Occupation Psychological distress Mental health Neuroticism Cognitive Markers	
Christensen *et al.* 2021	Poisson regression	Susceptibility Mortality	No association found	
Clift *et al.* 2022	Logistic regression and Mendelian randomisation	Susceptibility Mortality	Smoking	
Dabbah *et al.* 2021	Machine learning	Mortality	Kidney function Respiratory function	Pneumonia
Didikoglu *et al.* 2021	Logistic regression	Susceptibility Hospitalisation	Maternal smoking Breastfeeding Birthweight	
Elliott *et al.* 2021	Logistic regression	Mortality	Age Ethnicity Occupation Smoking Steroid use Cystatin C	Cardiovascular disease Hypertension Diabetes Autoimmune disease
Fan *et al.* 2020	Logistic regression	Susceptibility Mortality	Alcohol consumption	

Fatima *et al.* 2021	Logistic regression	Susceptibility	Occupation Ethnicity	
Freuer, Linseisen, and Meisinger 2021	Radial regression	Susceptibility Hospitalisation	BMI Waist circumference	
Gao *et al.* 2022	Cox regression	Susceptibility Hospitalisation Mortality	BMI Waist circumference	Type 2 diabetes
Gillies *et al.* 2022	Logistic regression	Hospitalisation Mortality	Living situation	
Hamer *et al.* 2020	Linear regression	Hospitalisation	Physical activity BMI C-reactive Protein	
Hassan *et al.* 2021	Logistic regression	Hospitalisation Mortality		Schizophrenia Bipolar disorder Depression
Hastie, Pell, and Sattar 2021	Cox regression	Hospitalisation Mortality	No association found	
Henne *et al.* 2021	Logistic regression		No association found	

Ho, Petermann-Rocha, *et al.* 2020	Poisson regression	Mortality	Age	
Ho, Celis-Morales, *et al.* 2020	Poisson regression	Hospitalisation	BMI Glycated haemoglobin Smoking Walking pace Medication Forced expiratory volume HDL cholesterol Ethnicity Deprivation Cystatin C	Hypertension
Hu *et al.* 2022	Logistic regression	Hospitalisation Mortality		Neurodegenerative disease
Q.-M. Huang *et al.* 2022	Logistic regression	Hospitalisation Mortality		COPD
Julkunen *et al.* 2021	Logistic regression	Hospitalisation		
Kolin *et al.* 2020	Poisson regression	Hospitalisation	Ethnicity Deprivation Medication Blood types	COPD Ischemic heart disease Mental disorders
Kuo *et al.* 2021	Logistic regression	Hospitalisation Mortality	Phenotypic Age	

Larvin *et al.* 2021	Logistic regression and cox	Hospitalisation Mortality	BMI	Periodontal disease
Harriet Larvin *et al.* 2020	Logistic regression	Hospitalisation Mortality		Periodontal disease
Lassale *et al.* 2020	Logistic regression	Hospitalisation	Ethnicity	
Lassale *et al.* 2021	Logistic regression	Hospitalisation Mortality	HDL cholesterol	
Lee *et al.* 2021	Logistic regression	Susceptibility	Deprivation Ethnicity BMI Occupation Smoking	Cancer
Lehrer and Rheinstein 2021b	ANOVA	Mortality	Medications	
Lehrer and Rheinstein 2021a	Logistic regression	Susceptibility	Eyewear	
X. Li *et al.* 2021	Logistic regression	Susceptibility Hospitalisation Mortality	No association	

	and Cox model			
J. Li *et al.* 2022	Logistic regression	Hospitalisation Mortality	Waist circumference Hip circumference	Non-alcoholic fatty liver disease
S. Li *et al.* 2021	Logistic regression	Susceptibility Hospitalisation	Vitamin D MUHO	
H. Li *et al.* 2021	Cox model	Mortality		Cancer
Liu *et al.* 2022	Logistic regression	Susceptibility Hospitalisation	Sleeping habits	
Lodge *et al.* 2021	Logistic regression	Susceptibility		Asthma
H. Ma *et al.* 2021	Logistic regression	Susceptibility	Vitamin D supplement use	
Y. Ma *et al.* 2021	Significance tests	Hospitalisation Mortality	Medication	
Maidstone *et al.* 2021	Logistic regression	Susceptibility	Occupation	
McQueenie *et al.* 2020	Poisson regression	Susceptibility	Medications	Multiple comorbidities

Papadopoulou *et al.* 2021	Phenome-wide association studies	Susceptibility Hospitalisation Mortality		Cardiovascular diseases
Patel *et al.* 2020)	Observational	Hospitalisation	Ethnicity Deprivation	
Petermann-Rocha *et al.* 2020	Poisson regression	Hospitalisation Mortality	Frailty score	
Peters, MacMahon, and Woodward 2021	Cox regression	Mortality	BMI Waist circumference Waist to hip ratio Waist to heigh ratio	
Prats-Uribe *et al.* 2021	Poisson regression	Susceptibility Mortality	Smoking	
Razieh *et al.* 2020	Logistic regression	Susceptibility	BMI Ethnicity	
Safizadeh *et al.* 2021	Logistic regression	Hospitalisation Mortality	Medications	

Sattar *et al.* 2020	Poisson regression	Susceptibility Hospitalisation Mortality	BMI Ethnicity	
Travaglio *et al.* 2021	Logistic regression	Susceptibility Mortality	PM and NOx air pollution	
Wang *et al.* 2021	Logistic regression	Susceptibility Mortality		Psychiatric disorders
Wong *et al.* 2021	Machine learning	Hospitalisation	Age Medication Waist circumference Kidney function	Type 2 diabetes Coronary artery disease Atrial fibrillation Dementia
Woodward, Peters, and Harris 2021	Cox regression	Mortality	Deprivation	
Xiang, Wong, and So 2021	Logistic regression	Susceptibility Hospitalisation Mortality	Medications	
Yang *et al.* 2020	Logistic regression	Susceptibility		Psychiatric disorders
Yates *et al.* 2020	Logistic regression	Susceptibility	BMI Waist circumference	

Yates *et al.* 2021	Logistic regression	Hospitalisation Mortality	BMI Walking pace	
Yoshikawa, Asaba, and Nakayama 2021	GWAS summary statistics	Susceptibility Hospitalisation	Triglyceride	
Raisi-Estabragh *et al.* 2020	Logistic regression	Hospitalisation	Ethnicity	
Zhang *et al.* 2021	Logistic regression	Susceptibility Hospitalisation	Blood group	
J. Zhou *et al.* 2021	Logistic regression	Hospitalisation		Neurodegenerative disease
Zhu *et al.* 2020a	Logistic regression	Hospitalisation		Asthma
Zhu *et al.* 2020b	Logistic regression	Hospitalisation	BMI Waist circumference	

Investigating the wide clinical spectrum and the causes of the manifold symptoms of COVID-19 requires a comprehensive view of the intricate and interwoven pathways of human biology. Diseases are often the result of complex causal chains, where the effect

of one risk factor can only be fully understood in the context of all others (Kraemer *et al.* 2001). All 66 UKB COVID-19 risk factor studies listed in Table 4-1 are based on the same cohort, yet few studies agree on all independent risk factors and even fewer draw from all available measurements. Commonly agreed-on risk factors also found in UKB studies like sex, age, obesity, and pulmonary and cardiovascular diseases were discovered early in the COVID-19 outbreak (Docherty et al. 2020; N. Chen et al. 2020; Huang et al. 2020; F. Zhou et al. 2020). Beyond these risk factors agreed upon shortly after the onset of the pandemic, multiple comorbidities and clinical measures are listed depending on the focus of the study. The identification of risk factors and the analysis of their independence hinges on which other variables are included in the model (Brotman *et al.* 2005). Focusing on one single aspect of human health, such as obesity, diabetes or periodontal disease ignores how the many factors of health interact, and potentially neglects unknown colliders that might actually be causal (Griffith *et al.* 2020). Using domain knowledge to select variables for study design can substantially reduce computation time but can also bias risk factor discovery. Variable selection in statistical inference is a subjective art that makes unstated assumptions—which may miss signals or reach premature conclusions—unless all alternative models are considered. In this chapter I aim to provide an agnostic approach by incorporating a vast array of information contained in UKB to fully leverage its potential for broadening the understanding of COVID-19 risk.

In the age of big data analysis, a popular method capable of including large quantities of heterogeneous data is machine learning. In contrast to most COVID-19 studies of UKB data which used classical statistical methods like logistic regression, Poisson regression

and Cox models, there were also two studies which used machine learning in the form of gradient boosted trees (Dabbah et al. 2021; K. C.-Y. Wong et al. 2021). Albeit primarily used for predicting outcome, machine learning generates relative feature importances as a by-product of their training (Saarela and Jauhiainen 2021). Machine learning can leverage correlation for prediction but as explained in the introduction, causality is never established and therefore without further investigation interpreting feature importance remains precarious (Prosperi *et al.* 2020). Machine learners are optimised for prediction and can be capable hypothesis generation (Khoury and Ioannidis 2014) but using gradient boosted trees alone will not help to establish causality. However, it does allow me to compare feature importance to inferential modelling to interrogate the hypotheses put forward by machine learning. Additionally, I can provide a prediction of whether individuals are likely to have severe forms of disease upon infection with SARS-CoV-2. I hope to thus give a broad view of what causes and who is at risk of severe COVID-19.

An alternative bottom-up risk factor discovery can be achieved within a Bayesian framework. The influence of variable choice on the outcome can be explicitly addressed via Bayesian model averaging (BMA) (Hinne *et al.* 2020). BMA provides a parameter estimate by averaging across multiple different models weighted by their respective probability (Hinne *et al.* 2020). In this context, the models correspond to different sets of potential risk factors. Here I use a Markov chain Monte Carlo (MCMC) algorithm to randomly sample from most available UKB variables for inclusion into a logistic regression model predicting different COVID-19 phenotypes. My MCMC chain start points are sampled broadly from all available variables, and through a stochastic

sampling process BMA generates the posterior inclusion probability (PoP). The PoP enables me to estimate the effect of all variables on COVID-19 outcome whilst accounting for model uncertainty. To speed up computations, the method uses a Normal distribution approximation for the likelihood function, a simplification that is motivated by the large sample sizes analysed. This chapter compares BMA and gradient boosted trees to investigate COVID-19 severity, hospitalisation, and susceptibility. I offer a broad view on risk that harvests the strengths of machine learning for prediction and Bayesian modelling for inference. The agnostic and comprehensive risk factor assessment through BMA sheds light on which reported risk factors are likely to be robust to model misspecification and allows me to investigate how feature importance compares to inferential modelling. I hope to thus offer a comprehensive data-driven analysis of COVID-19 risk based on UKB data that can be readily repurposed for other diseases.

4.2 Methods

4.2.1 Dataset Collection and Preparation

Drawing robust inferences from risk factor analysis is dependent on the quality of the underlying data, which requires rigorous data cleaning. The following steps were taken to clean up UK Biobank data. Participants that died or were otherwise lost to follow up since the initial UK Biobank registration were excluded. As some COVID-19 test results were only provided for England, non-English participants were also excluded. Variables missing more than 5% of the data were excluded and columns were typecast according to UK Biobank data specifications. Integer and continuous missing values were imputed using the mean of the column, and 3 columns with zero variance were removed. Missing factors were cast as a new level and encoded as the integer -999. Factors with more

than 50 levels were removed, excluding 2 columns. Numerous biomarkers have multiple measurements available for each participant, the result of one or more hospital visits after initial UK Biobank registration. For such repeated factor measurements, the median of all values was computed, and averages over all visits were used for continuous values. All factors with more than 2 levels were dummy-encoded as one level versus all other levels. Diseases were encoded in the form proposed by the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD) (World Health Organization 1992). The ICD codes were aggregated in Charlson comorbidities indices, which groups similar diseases together (Charlson *et al.* 1987). The final dataset comprised of 426,893 rows representing participants and 1,143 columns consisting of measurements, of which 155 were continuous variables, 74 integer values and 914 binary categorical dummy variables. The dummy variables were the result of the one-versus-rest encoding of 17 Charlson codes, 341 ICD codes and 141 other factor columns. With the data processing finished, the phenotypes of risk were defined as follows.

4.2.2 Phenotype Definitions

The risk assessment was formalised as a binary classification task with case and control definitions drawn from the COVID-19 Host Genetics Initiative (The Severe COVID-19 GWAS Group 2020):

- **Very severe:** Cases were defined as all patients hospitalised, dead or on respiratory support with laboratory-confirmed SARS-CoV-2 infection and all patients admitted to hospitals with COVID-19 listed as primary reason for admission (n = 871). Controls were all residual UKB participants in the final dataset (n = 426,022).

- **Hospitalisation:** Cases were defined as all hospitalised, laboratory-confirmed SARS-CoV-2 infections that were hospitalised due to COVID-19 symptoms (n = 2,681). Controls were all residual UKB participants in the final dataset (n = 424,262).
- **Susceptibility:** Cases were defined by laboratory-confirmed SARS-CoV-2 and all COVID-19 infections confirmed by medical professionals (n = 77,867). Controls were all residual UKB participants in the final dataset (n = 349,026).

The phenotypes were subsequently used as classes for risk prediction in the machine learning model and as outcomes for the logistic regression in the BMA approach.

4.2.3 Machine Learning

4.2.3.1 Classifier Choice and Training

For machine learning risk prediction, gradient boosted decision trees were used because they generate interpretable and reliable classification models, whilst requiring little data pre-processing. The adequacy of this algorithm for COVID-19 UKB studies has already been demonstrated twice (K. C.-Y. Wong et al. 2021; Dabbah et al. 2021). LightGBM (Ke *et al.* 2017) was chosen as an implementation due to it natively supporting factors which are abundant in the UK Biobank. The model was trained on a75% training set and the performance was measured on a 25% test set with samples being randomly sorted into either set. Log loss was chosen as an optimisation function which generates an output score that ranges from 0 to 1, resembling probability. The trees were grown using loss guide as a growth policy. The residual parameters of the model were estimated through hyper-parameter tuning using Python.

In choosing optimal hyperparameters, the training set was further split into a training set (90%) and an evaluation set (10%) so the latter could be used for early stopping to combat overfitting. A greedy linear approach was used to arrive at optimal parameters, whereby a succession of hyper-parameters was established. For every parameter a pre-defined range of parameters was cycled through by using 5-fold cross-validation to assess performance. Among all possible values for one hyperparameter, the value resulting in the lowest log loss was kept for tuning the next hyperparameter. The training was stopped when the log loss on the evaluation set did not improve for 10 consecutive iterations. The trained classifier was applied on the training and testing set with performance measured in precision, recall, AUC, F1, negative predictive value and accuracy (see Figure 1-3 for an explanation of performance metrics). As class balance is important for machine learning validation, the test set was balanced by randomly undersampling the controls to the number of cases in the dataset. For comparison with the PoP as estimated by BMA, relative feature importance was measured in several different measures.

4.2.3.2 Feature Importance

Relative feature importance was measured by using the “gain” and “split” importances implemented in LightGBM in addition to Shapley values (Huettner and Sunder 2012) and permutation importances. The SHAP package (available at: https://shap.readthedocs.io /en/latest/api.html) was used for generating Shapley values. Permutation importances were measured by iterating through all columns of the data and randomly shuffling one column at a time. The machine learner was trained on the complete data with one column shuffled. The prediction of the shuffled column classifier was compared against

a baseline prediction from a classifier using the original unshuffled dataset. The predictions were compared using log loss, which was used as an approximation of feature importance for the shuffled column. To better judge which features were used by the classifier, random noise, Gaussian noise, Poisson noise and a random categorical column were added to the dataset as all available noise distributions in the numpy Python library. Features scoring lower than any of the noise columns were considered irrelevant and their importance was set to 0. All feature comparisons were correlated against the posterior probability computed by the BMA using Spearman's rank coefficient. All machine learning feature importances were scaled so that the maximum value was equal to 1 and the minimum value was equal to 0. The comparison of feature importances to PoP should establish how well feature importance (as a by-product of prediction) approximates the inferences made by the following BMA approach.

4.2.4 Bayesian Model Averaging[1]

4.2.4.1 Notation

A general regression setting was used in which there are n observed outcomes $Y_1, \ldots, Y_n$ and two groups of parameters. The first group of v parameters, $\beta_1, \ldots, \beta_v$, are regression coefficients for v candidate explanatory variables (i.e. features), X_{ij}, $i = 1 \ldots n$, $j = 1 \ldots v$. The primary aim was to identify which candidate variables influence the outcome, i.e. which of the regression coefficients are non-zero. The second group of ζ parameters, $\gamma_1, \ldots, \gamma_\zeta$, were of secondary interest due to their inclusion in every iteration of the model. The ζ parameters could include an intercept term, regression coefficients

[1] Except for 4.2.5.3, these sections were written by Daniel J. Wilson who also formulated the theory. Nicolas Arning rewrote the section and contributed subsection 4.2.5.3 and the theory behind it.

for other variables that are always included in the model, and an error variance, among others.

In total there were 2^{v} models which were indexed by ***s***, a binary vector of length *v* in which the *j*th element indicates the inclusion of the *j*th candidate explanatory variable, i.e.

$$s_j = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0. \end{cases} \quad (1)$$

The notation $\mathcal{M}_{\boldsymbol{s}}$ refers to the model with the constraints specified by **s**. The notation $\mathcal{M}_{0}$ refers to the 'grand null 'in which $s_j = 0$ for all $j = 1 \ldots v$.

Results were summarised with respect to one or more specific variable(s)-of-interest. The number of variables-of-interest was denoted as $v_{\oplus}$, with $1 \leq v_{\oplus} \leq v$, where commonly the focus lay on $v_{\oplus} = 1$. When considering these variables-of-interest, the $2^{v-v_{\oplus}}$ combinations of other variables were considered as ordered in a way where $\mathcal{M}_{i\oplus}$ was used to represent a model including the variable(s)-of-interest and $\mathcal{M}_{\ominus}$ a model excluding the variable(s)-of-interest, with $1 \leq i \leq 2^{v-v_{\oplus}}$.

4.2.4.2 Targets of Inference

There were two main targets of inference:

- Estimating the posterior odds that one or more specific variable(s)-of-interest were associated with the outcome

$$\mathrm{PoO}_{\oplus} = \frac{\Pr(\mathcal{M}_{\oplus} \mid \boldsymbol{Y}, \boldsymbol{X})}{\Pr(\mathcal{M}_{\ominus} \mid \boldsymbol{Y}, \boldsymbol{X})},$$

  where $\mathcal{M}_{\oplus}$ and $\mathcal{M}_{\ominus}$ are shorthand for all models that respectively include and exclude the variable(s)-of-interest, allowing the above to be rewritten as

$$\mathrm{PoO}_{\oplus} = \frac{\sum_{i=1}^{2^{\nu-\nu_{\oplus}}} \Pr(\mathcal{M}_{i\oplus} \mid \boldsymbol{Y}, \boldsymbol{X})}{\sum_{i=1}^{2^{\nu-\nu_{\oplus}}} \Pr(\mathcal{M}_{i\ominus} \mid \boldsymbol{Y}, \boldsymbol{X})}$$

- Estimating the magnitude and direction of effect of specific variable(s)-of-interest on the outcome, averaged over all models in which those variables were included. Specifically, the posterior mean and variance

$$\mathbb{E}\left[\boldsymbol{\beta}_{\oplus} \mid \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}_{\oplus}\right] \quad \text{and} \quad \mathbb{V}\left[\boldsymbol{\beta}_{\oplus} \mid \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}_{\oplus}\right]$$

  were estimated.

In the above, ***Y*** represents the outcome data and ***X*** represents all candidate variables (i.e. features) potentially associated with the outcome. The targets of inference were estimated by MCMC.

4.2.4.3 Initialising the MCMC

We originally explored a strategy in which the MCMC was initialised by choosing 10 candidate variables uniformly at random. However, a strong correlation structure between candidate variables made this an inefficient strategy. Instead, a furthest neighbour approach was used to optimally explore the feature space for chain initialisation. Spearman's rank correlation between all variables was calculated and all pairs of variables above 0.5 correlation were collected, which was chosen manually from a distribution of all observed correlations. A random variable in the largest group was chosen as a starting point of a chain. The remaining variables for initialisation were selected by iteratively choosing the variable least correlated to the variables already included. If there were multiple variables exhibiting the minimum observed correlation, for example 0, one was chosen at random. Initially the variable was chosen which was the least correlated with the starting variable. When selecting the third and later variables, "least correlated" was decided by computing the average distance from the

already selected variables to the remaining ones. The procedure was repeated until a starting group of 10 variables was assembled. Using a furthest neighbour approach for starting MCMC chains led to a better sampling of the feature space contained in the UKB than an entirely random chain initialisation.

4.2.4.4 Metropolis-Hastings Moves to Explore the Model Space

Updates to the model were proposed at each step by adding or removing one variable at a time or swapping out one variable for another correlated variable. In the notation, $\boldsymbol{s}$ denotes the variable inclusion vector explored by iteration t of the MCMC, and $|\boldsymbol{s}| = \sum_{i=1}^{v} s_i$ shows the number of variables the vector included. The probability with which each move was proposed is summarised in Table 4-2. The parameter μ was determined by the prior distribution on the number of explanatory variables associated with the outcome (see below).

Table 4-2 Metropolis-Hastings Moves

	When $\|\boldsymbol{s}\| = 0$	When $0 < \|\boldsymbol{s}\| < v$	When $\|\boldsymbol{s}\| = v$
Add a variable	1	$\frac{19(1-\mu)}{20(2-\mu)}$	0
Remove a variable	0	$\frac{19}{20(2-\mu)}$	1
Swap a variable	0	$\frac{1}{20}$	0

4.2.4.4.1 Adding a variable

A variable was added one at a time by uniform random choice from the $v - |\boldsymbol{s}|$ excluded variables, to create a new variable inclusion vector $\boldsymbol{s}'$.

4.2.4.4.2 Removing a variable

A variable was removed one at a time by uniform random choice from the $|\boldsymbol{s}|$ included variables, to create a new variable inclusion vector $\boldsymbol{s}'$.

4.2.4.4.3 Swapping a variable

A variable was chosen uniformly at random to be swapped out from among the $|\boldsymbol{s}|$ included variables. A replacement variable was chosen from among the $v - |\boldsymbol{s}|$ excluded variables with probability proportional to its squared correlation coefficient with the variable chosen for removal.

4.2.4.4.4 Acceptance probabilities

From the usual theory for Metropolis-Hastings moves (Metropolis *et al.* 1953), taking into account the likelihood $\Pr(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_s)$ (specified below), prior model probability $\Pr(\mathcal{M}_s)$ (specified below) and Hastings ratio, the proposal was accepted with probability

$$\min\left\{1, \frac{\Pr(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_{\boldsymbol{s}'})}{\Pr(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_{\boldsymbol{s}})} \frac{\Pr(\mathcal{M}_{\boldsymbol{s}'})}{\Pr(\mathcal{M}_{\boldsymbol{s}})} \frac{K(\boldsymbol{s}' \to \boldsymbol{s})}{K(\boldsymbol{s} \to \boldsymbol{s}')}\right\} \quad (2)$$

where the proposal probabilities in the Hastings ratio were:

$$K(\boldsymbol{s} \to \boldsymbol{s}') = \begin{cases} \frac{1}{\nu} & \text{if } |\boldsymbol{s}| = 0 \text{ and } \boldsymbol{s}' \text{ added a variable to } \boldsymbol{s} \\ \frac{19}{20}\frac{(1-\mu)}{(2-\mu)}\frac{1}{\nu - |\boldsymbol{s}|} & \text{if } 0 < |\boldsymbol{s}| < \nu \text{ and } \boldsymbol{s}' \text{ added a variable to } \boldsymbol{s} \\ \frac{1}{20}\frac{1}{|\boldsymbol{s}|}\frac{(1-s_j)\, r_{ij}^2}{\sum_{k=1}^{\nu}(1-s_k)\, r_{ik}^2} & \text{if } 0 < |\boldsymbol{s}| < \nu \text{ and } \boldsymbol{s}' \text{ replaced variable } i \text{ in } \boldsymbol{s} \text{ with } j \\ \frac{19}{20}\frac{1}{(2-\mu)}\frac{1}{|\boldsymbol{s}|} & \text{if } 0 < |\boldsymbol{s}| < \nu \text{ and } \boldsymbol{s}' \text{ removed a variable from } \boldsymbol{s} \\ \frac{1}{\nu} & \text{if } |\boldsymbol{s}| = \nu \text{ and } \boldsymbol{s}' \text{ removed a variable from } \boldsymbol{s} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

with $|\boldsymbol{s}| = \sum_{i=1}^{\nu} s_i$ and r_{ij}^2 the squared correlation coefficient between candidate variables *i* and *j*.

4.2.4.5 *Summarising posterior odds ratios*

For every phenotype definition, i.e. very severe COVID-19, hospitalisation, and susceptibility, 50 chains of length 100,000 iterations were run. A burn-in period of 1000 iterations was removed, deemed sufficient based on preliminary runs.

The posterior probability that candidate variable *i* was associated with the outcome was defined as

$$\widehat{\mathrm{PoP}}_{i\oplus} = \frac{1}{N_{\mathrm{iter}}} \sum_{t=1}^{N_{\mathrm{iter}}} s_i^{(t)} \quad (4)$$

where N_{iter} was the total number of iterations, and $\boldsymbol{s}^{(t)}$ was the inclusion vector during iteration *t* of the MCMC, excluding burn-in.

A standard error for $\widehat{PoP}$ was estimated by comparing the results across the independent chains:

$$\mathrm{s.e.}\left(\widehat{\mathrm{PoP}}_{i\oplus}\right) = \sqrt{\frac{\sum_{c=1}^{N_{\mathrm{chains}}} \left(\widehat{\mathrm{PoP}}_{i\oplus}^{(c)} - \widehat{\mathrm{PoP}}_{i\oplus}\right)^2}{N_{\mathrm{chains}}}} \quad (5)$$

where N_{chains} was the total number of chains, and $\widehat{PoP}^{(c)}$ was the estimated posterior odds from chain *c* alone. The standard error enabled the assessment of an approximate 95% confidence interval for the posterior probabilities as

$$\widehat{\mathrm{PoP}}_{i\oplus} \pm 2 \text{ s.e.} \left(\widehat{\mathrm{PoP}}_{i\oplus}\right) \tag{6}$$

The confidence intervals were used to assess the level of stochastic error due to MCMC sampling, and thereby also judge the length of chains needed. The posterior probabilities and their approximate 95% confidence intervals were converted into posterior odds using the formula

$$\mathrm{PoO} = \frac{\mathrm{PoP}}{1 - \mathrm{PoP}} \tag{7}$$

An explanation of the prior and likelihood functions used in the MCMC follows.

4.2.4.6 Prior model probabilities

A priori, every candidate explanatory variable was assumed to be equally likely to be associated with the outcome.

For the number of candidate variables included in the model, a truncated geometric distribution on 0, 1, . . ., *v* with parameter $\mu = 0.1$ was assumed:

$$\Pr(|\boldsymbol{s}| = x) = \frac{\mu\,(1-\mu)^x}{1-(1-\mu)^{\nu+1}} \tag{8}$$

The prior expectation on the number of included variables under this prior, for large *v*, is approximately $(1-\mu)/\mu$. A μ of 0.1 was chosen, implying an expectation of about 9 variables associated with the outcome based on preliminary runs of the algorithm.

This prior is over-dispersed relative to a binomial distribution with the same prior expectation, which was preferred on the grounds that it was therefore less informative.

The prior probability on the model with inclusion vector ***s*** was thus

$$\Pr(\mathcal{M}_{\boldsymbol{s}}) = \frac{\mu\,(1-\mu)^{|\boldsymbol{s}|}}{1-(1-\mu)^{\nu+1}} \Bigg/ \left(\frac{\nu!}{|\boldsymbol{s}|!\,(\nu-|\boldsymbol{s}|)!}\right) \tag{9}$$

4.2.4.7 Full likelihood for model and parameters

For a given inclusion vector ***s***, the association between explanatory variables and outcome was analysed using a linear model. The outcome variables of interest were binary, with $Y_i = 1$ indicating a case, and $Y_i = 0$, a control. Therefore, a logistic regression with linear predictor was assumed

$$\text{logit}\left(\Pr(Y_i = 1|\boldsymbol{X}, \mathcal{M}_{\boldsymbol{s}}, \boldsymbol{\beta}, \boldsymbol{\gamma})\right) = \sum_{j=1}^{\nu} \beta_j\, X_{ij} + \gamma, \tag{10}$$

where β_j is the effect of explanatory variable j, constrained to be zero when $s_j = 0$, and γ is the intercept term, included in every model. Here the logit function is $\text{logit}(x) = \log(x/(1-x))$. The n observations were assumed independent.

However, the parameters β and γ were not explored using the MCMC, since this would have been inefficient. Instead, the parameters were integrated over analytically, with the help of a large-sample Normal approximation to the likelihood defined above.

Henceforth $\theta = \begin{pmatrix}\beta\\ \gamma\end{pmatrix}$ denotes the combined parameter vector. It was important to adopt a precise notation for the elements of θ that are free in a specific model as it is a potential source of confusion. I use $\boldsymbol{F_s}$ to denote the index set of free (i.e. unconstrained)

parameters in model $\mathcal{M}_s$. Similarly, β_s = $\boldsymbol{F_s} \setminus \boldsymbol{F_0}$ denotes the index set of parameters that are free in model $\mathcal{M}_s$,but not in $\mathcal{M}_0$.

A large sample size approximation (see subsection 2.4.11 later for mathematical details) allows the likelihood to be written as

$$\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_{\boldsymbol{s}},\theta\right) \sim \begin{cases} c_{\boldsymbol{s}}\ f_{\text{Normal}}\left(\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}} \,\middle|\, \theta_{\mathcal{F}_{\boldsymbol{s}}}, \{n\ \mathcal{I}(\mathbf{0})_{\mathcal{F}_{\boldsymbol{s}},\mathcal{F}_{\boldsymbol{s}}}\}^{-1}\right) & \text{if } \theta \in \Theta_{\mathcal{M}_{\boldsymbol{s}}} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Here c_s is a constant, f_{Normal} is the multivariate Normal density function and $\hat{\theta}^{(S)} = \begin{pmatrix}\hat{\beta}^{(S)} \\ \hat{\gamma}^{(S)}\end{pmatrix}$ is the maximum likelihood estimate (MLE) of θ under model $\mathcal{M}_s$. Here $\Theta_{\mathcal{M}_S}$ is the parameter space of model $\mathcal{M}_s$, which imposes the constraint that $\beta_i = 0$ when $s_i = 0$, and

$$\mathcal{I}(\theta) = -\frac{1}{n}\nabla\nabla'\sum_{\boldsymbol{y}}\left[\log\Pr(\boldsymbol{y}|\boldsymbol{X},\theta)\right]\Pr(\boldsymbol{y}|\boldsymbol{X},\theta) \tag{12}$$

is the per-observation Fisher information matrix (FIM).

This form of the likelihood enables the analytical integration over the prior distribution of θ, described next.

4.2.4.8 Prior distribution of effects

It is convenient to assume a Normal prior distribution for the free parameters, as it is conjugate to the approximate Normal likelihood. The assumption simplifies computations and involves only two parameters, a mean and variance.

A mean of zero was chosen giving positive and negative effects equal weight and centring the distribution of effects under the alternative hypothesis (that a candidate

variable influences outcome) around the null hypotheses (that a candidate variable has zero effect on outcome).

A pragmatic approach was taken for the variance, assuming a form that would give convenient results. Therefore, *a priori*

$$\theta_{\mathcal{F}_s} \mid \boldsymbol{X}, \mathcal{M}_s \sim N\left(\mathbf{0}, \{h\, \mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\}^{-1}\right) \tag{13}$$

was assumed where *h* is a hyper-parameter controlling the dispersion of the prior.

This prior is essentially Zellner's (1986) g-prior applied to an approximate Normal likelihood. Its use produces a convenient form for the model likelihood integrated over the effect sizes, which is needed for the MCMC, and explained next.

4.2.4.9 Model likelihood integrated over effect sizes

To streamline the MCMC, θ was integrated over analytically, which was possible because of the mathematically convenient forms provided by assuming an asymptotically Normal likelihood, and a conjugate Normal prior distribution on θ.

The analytical integration of θ allowed for the defining the likelihood of $\mathcal{M}_s$ in relation to $\mathcal{M}_0$ (see section 2.4.11 for mathematical details); this is known as the Bayes factor:

$$\begin{aligned} \mathrm{BF}_s &= \frac{\Pr(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M}_s)}{\Pr(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M}_0)} \\ &= \left(\frac{h}{n+h}\right)^{|s|/2} R_s{}^{n/(n+h)} \end{aligned} \tag{14}$$

where

$$R_s = \frac{\Pr\left(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M}_s, \hat{\theta}^{(s)}\right)}{\Pr\left(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M}_0, \hat{\theta}^{(0)}\right)} \tag{15}$$

is the maximised likelihood ratio. This Bayes factor was sufficient for use in the MCMC, because the acceptance probability (Equation 2) always depends on a ratio of model likelihoods for $\mathcal{M}_s$ versus $\mathcal{M}_{s'}$.

The strength of this form lies in not depending on an explicit computation of the FIM. The only quantities needed were:

- The maximised likelihood from the `glm` command in R
- Knowledge of the sample size, *n*
- A choice of the prior hyper-parameter, *h*

For *n* >> *h*, the main role of the hyper-parameter *h* was to modify the penalty on the number of free parameters in the Bayesian interpretation of the maximised likelihood ratio R_s.

It was natural under the defined prior to assume that $h \sim O(1)$, because this implies that the weight of evidence provided by the prior was comparable to that from a single observed datapoint. Therefore, *h* was usually set to 1, but the robustness of the final results to a weaker prior of $h = 0.1$ was also investigated.

4.2.4.10 Simulating effect sizes

The second target-of-inference was the effect size for the explanatory variables, conditional on inclusion in the model. Here, direct simulation was possible, because the conditional posterior distribution for the free parameters in the model had the Normal distribution

$$\theta_{\mathcal{F}_s} \,|\, \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}_s \;\sim\; N\left(\frac{n}{n+h}\hat{\theta}^{(s)}_{\mathcal{F}_s}, \frac{n}{n+h}\left(n\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\right)^{-1} \right) \tag{16}$$

For each iteration of the MCMC, the effect sizes were directly simulated from this conditional posterior distribution, using the current state of $\mathcal{M}_s$. The simulations were done on a variable-by-variable basis (i.e. $v_{\oplus}$ = 1) using the R function `rnorm`. This required the classical maximum likelihood estimate $\widehat{\theta}_{Fs}^{(s)}$ and variance ($n\ I(\boldsymbol{O})_{Fs,Fs}$)$^{-1}$. Both were obtained by applying `coef(summary(fit))` to the object `fit` output by the `glm` command, which had already been run during the MCMC to obtain the maximised likelihood.

Finally, the following summaries for variable-of-interest *i* were computed:

$$\hat{\mathbb{E}}\left[\boldsymbol{\beta}_{\oplus} \mid \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}_{\oplus}\right] = \frac{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)}\, \beta_i^{(t)}}{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)}} \tag{17}$$

and

$$\hat{\mathbb{V}}\left[\boldsymbol{\beta}_i \mid \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}_{\oplus}\right] = \frac{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)} \left(\beta_i^{(t)}\right)^2}{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)}} - \left(\frac{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)}\, \beta_i^{(t)}}{\sum_{t=1}^{N_{\text{iter}}} s_i^{(t)}}\right)^2 \tag{18}$$

where $\beta_i^{(t)}$ was the effect size simulated for variable *i* during iteration *t* of the MCMC. Note that it was not necessary to simulate$\beta_i^{(t)}$ for iterations when the variable was excluded from the model ($s_i^{(t)} = 0$); which would have been constrained to zero by definition regardless.

4.2.4.11 Mathematical details

Some of the more technical details underpinning the above are described below.

2.4.11.1 Approximate likelihood

In calculating an approximate likelihood (section 2.4.7), the sample size *n* was assumed to be large, so that the likelihood surface could be approximated via a second-order Taylor series expansion as a Normal distribution (Cox and Hinkley 1974), i.e.

$$\begin{aligned}\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_s,\theta\right) &\sim \begin{cases} \Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_s,\hat{\theta}^{(s)}_{\mathcal{F}_s}\right)\times \\ \quad \exp\left\{-\frac{n}{2}\left(\theta_{\mathcal{F}_s}-\hat{\theta}^{(s)}_{\mathcal{F}_s}\right)'\hat{\mathcal{I}}(\hat{\theta}^{(s)})_{\mathcal{F}_s,\mathcal{F}_s}\left(\theta_{\mathcal{F}_s}-\hat{\theta}^{(s)}_{\mathcal{F}_s}\right)\right\} & \text{if } \theta\in\Theta_{\mathcal{M}_s} \\ 0 & \text{otherwise}\end{cases} \\ &= \begin{cases} c_s\, f_{\text{Normal}}\left(\hat{\theta}^{(s)}_{\mathcal{F}_s}\middle|\,\theta_{\mathcal{F}_s},\left\{n\,\hat{\mathcal{I}}(\hat{\theta}^{(s)})_{\mathcal{F}_s,\mathcal{F}_s}\right\}^{-1}\right) & \text{if } \theta\in\Theta_{\mathcal{M}_s} \\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{19}$$

Here c_s is a constant, f_{Normal} is the multivariate Normal density function, $\hat{\theta}^{(S)} = \begin{pmatrix}\hat{\beta}^{(S)} \\ \hat{\gamma}^{(S)}\end{pmatrix}$ is the maximum likelihood estimate (MLE) of θ under model $\mathcal{M}_s$, and Θ_{M_S} is the parameter space of model M_s. This expansion imposes the constraint that $\beta_i = 0$ when $s_i = 0$, and

$$\hat{\mathcal{I}}(\hat{\theta}^{(s)})_{\mathcal{F}_s,\mathcal{F}_s} = -\frac{1}{n}\nabla\nabla'\log\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_s,\theta_{\mathcal{F}_s}\right)\Big|_{\theta_{\mathcal{F}_s}=\hat{\theta}^{(s)}_{\mathcal{F}_s}} \tag{20}$$

is the per-observation empirical FIM.

A second approximation allowed Equation 19 to be rewritten

$$\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_s,\theta\right) \sim \begin{cases} c_s\, f_{\text{Normal}}\left(\hat{\theta}^{(s)}_{\mathcal{F}_s}\middle|\,\theta_{\mathcal{F}_s},\{n\,\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\}^{-1}\right) & \text{if } \theta\in\Theta_{\mathcal{M}_s} \\ 0 & \text{otherwise}\end{cases} \tag{21}$$

where

$$\mathcal{I}(\theta) \quad = \quad -\frac{1}{n}\nabla\nabla' \sum_{\boldsymbol{y}} [\log \Pr(\boldsymbol{y}|\boldsymbol{X},\theta)] \Pr(\boldsymbol{y}|\boldsymbol{X},\theta) \tag{22}$$

is the per-observation FIM. This is justifiable because, assuming a large sample size n, and subject to important technical conditions (Cox and Hinkley 1974) assumed here but not expounded,

$$\mathcal{I}(\theta^{(\boldsymbol{s})}) \sim \hat{\mathcal{I}}(\hat{\theta}^{(\boldsymbol{s})}) \tag{23}$$

Intuitively, this is the asymptotic equivalence between the empirical FIM evaluated at the MLE, and the FIM evaluated at an arbitrary point. The equivalence is implied by (i) the asymptotically Normal log-likelihood surface, which has a fixed Hessian matrix and (ii) the asymptotically negligible stochastic noise in the estimate of the FIM that the empirical FIM

provides.

The MLE under model M_s is defined as

$$\hat{\theta}^{(\boldsymbol{s})} \quad = \quad \text{argmax}_{\theta \in \Theta_{\mathcal{M}_{\boldsymbol{s}}}} \Pr(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_{\boldsymbol{s}},\theta) \tag{24}$$

Assuming the model does not suffer from identifiability issues – in particular, a problem with collinearity between the included variables – then the MLE is efficient to compute using standard software. The `glm` function in R was used, whilst not allowing non-identifiable models by assuming they had zero likelihood. This meant that any non-identifiable model proposed during the MCMC was automatically rejected.

2.4.11.2 Model likelihood integrated over effect sizes

To streamline the MCMC, I integrated over θ analytically. The integration was possible because of the mathematically convenient forms provided by assuming (i) an asymptotically Normal likelihood, and (ii) a conjugate Normal prior distribution on θ. The model likelihood integrated over the prior could therefore be written as

$$\begin{aligned}
\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_{\boldsymbol{s}}\right) &= c_{\boldsymbol{s}} \int f_{\text{Normal}}\left(\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}} \middle| \theta^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}}, \{n\,\mathcal{I}(\mathbf{0})_{\mathcal{F}_{\boldsymbol{s}},\mathcal{F}_{\boldsymbol{s}}}\}^{-1}\right) \times \\
&\qquad f_{\text{Normal}}\left(\theta^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}} \middle| \mathbf{0}, \{h\,\mathcal{I}(\mathbf{0})_{\mathcal{F}_{\boldsymbol{s}},\mathcal{F}_{\boldsymbol{s}}}\}^{-1}\right) \mathrm{d}\theta^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}} \\
&= c_{\boldsymbol{s}}\, f_{\text{Normal}}\left(\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_{\boldsymbol{s}}} \middle| \mathbf{0}, \boldsymbol{\Sigma}\right) \qquad \text{with} \qquad \boldsymbol{\Sigma} = \left(\frac{1}{h}+\frac{1}{n}\right)\left(\mathcal{I}(\mathbf{0})_{\mathcal{F}_{\boldsymbol{s}},\mathcal{F}_{\boldsymbol{s}}}\right)^{-1} \qquad (25) \\
&= c_{\boldsymbol{s}}\, f_{\text{Normal}}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}} \middle| \mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\right) \times \\
&\qquad f_{\text{Normal}}\left(\hat{\boldsymbol{\gamma}}^{(\boldsymbol{s})} \middle| \boldsymbol{\Sigma}_{\mathcal{F}_0,\mathcal{B}_{\boldsymbol{s}}}\{\boldsymbol{\Sigma}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\}^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}}, \left[\{\boldsymbol{\Sigma}^{-1}\}_{\mathcal{F}_0,\mathcal{F}_0}\right]^{-1}\right), \qquad (26)
\end{aligned}$$

where the last line factorises the likelihood.

Similar steps can be followed to compute the grand null model likelihood, plugging in the appropriate change of prior. However, this requires factorisation before integrating:

$$\begin{aligned}
\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_0\right) &= c_{\boldsymbol{s}} \int f_{\text{Normal}}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}} \middle| \mathbf{0}, \boldsymbol{\Xi}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\right) \times \\
&\qquad f_{\text{Normal}}\left(\hat{\boldsymbol{\gamma}}^{(\boldsymbol{s})} \middle| \boldsymbol{\gamma} + \boldsymbol{\Xi}_{\mathcal{F}_0,\mathcal{B}_{\boldsymbol{s}}}\{\boldsymbol{\Xi}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\}^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}}, \left[\{\boldsymbol{\Xi}^{-1}\}_{\mathcal{F}_0,\mathcal{F}_0}\right]^{-1}\right) \times \\
&\qquad f_{\text{Normal}}\left(\boldsymbol{\gamma} \middle| \mathbf{0}, \{h\,\mathcal{I}(\mathbf{0})_{\mathcal{F}_0,\mathcal{F}_0}\}^{-1}\right) \mathrm{d}\boldsymbol{\gamma} \\
&\qquad \text{with} \qquad \boldsymbol{\Xi} = \frac{1}{n}\left(\mathcal{I}(\mathbf{0})_{\mathcal{F}_{\boldsymbol{s}},\mathcal{F}_{\boldsymbol{s}}}\right)^{-1} = \frac{h}{n+h}\boldsymbol{\Sigma}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}} \qquad (27) \\
&= c_{\boldsymbol{s}}\, f_{\text{Normal}}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}} \middle| \mathbf{0}, \frac{h}{n+h}\boldsymbol{\Sigma}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\right) \times \\
&\qquad f_{\text{Normal}}\left(\hat{\boldsymbol{\gamma}}^{(\boldsymbol{s})} \middle| \boldsymbol{\Sigma}_{\mathcal{F}_0,\mathcal{B}_{\boldsymbol{s}}}\{\boldsymbol{\Sigma}_{\mathcal{B}_{\boldsymbol{s}},\mathcal{B}_{\boldsymbol{s}}}\}^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_{\boldsymbol{s}}}, \left[\{\boldsymbol{\Sigma}^{-1}\}_{\mathcal{F}_0,\mathcal{F}_0}\right]^{-1}\right), \qquad (28)
\end{aligned}$$

From here, likelihood of $\mathcal{M}_s$ can be defined relative to $\mathcal{M}_0$; this is known as the Bayes factor:

$$\begin{aligned}
\mathrm{BF}_{\boldsymbol{s}} &= \frac{\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_{\boldsymbol{s}}\right)}{\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\mathcal{M}_0\right)} \\
&= \frac{f_{\text{Normal}}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\middle|\,\mathbf{0},\boldsymbol{\Sigma}_{\mathcal{B}_s,\mathcal{B}_s}\right)}{f_{\text{Normal}}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\middle|\,\mathbf{0},\frac{h}{n+h}\boldsymbol{\Sigma}_{\mathcal{B}_s,\mathcal{B}_s}\right)} \qquad (29) \\
&= \left(\frac{h}{n+h}\right)^{|\boldsymbol{s}|/2}\exp\left\{\frac{1}{2}\frac{n}{h}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right)'\left[\boldsymbol{\Sigma}_{\mathcal{B}_s,\mathcal{B}_s}\right]^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right\} \\
&= \left(\frac{h}{n+h}\right)^{|\boldsymbol{s}|/2}\exp\left\{\left(1-\frac{h}{n+h}\right)\frac{n}{2}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right)'\left[\left\{\left(\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\right)^{-1}\right\}_{\mathcal{B}_s,\mathcal{B}_s}\right]^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right\} \qquad (30) \\
&= \left(\frac{h}{n+h}\right)^{|\boldsymbol{s}|/2}{R_{\boldsymbol{s}}}^{n/(n+h)} \qquad (31)
\end{aligned}$$

The final step relies on a large-sample size identity for the maximised likelihood ratio between model $\mathcal{M}_s$ and the nested grand null $\mathcal{M}_0$:

$$\begin{aligned}
R_{\boldsymbol{s}} &= \frac{\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\hat{\theta}^{(\boldsymbol{s})}\right)}{\Pr\left(\boldsymbol{Y}|\boldsymbol{X},\hat{\theta}^{(0)}\right)} \qquad (32) \\
&= \exp\left\{\frac{n}{2}\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right)'\left[\left\{\left(\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\right)^{-1}\right\}_{\mathcal{B}_s,\mathcal{B}_s}\right]^{-1}\hat{\boldsymbol{\beta}}^{(\boldsymbol{s})}_{\mathcal{B}_s}\right\} \qquad (33)
\end{aligned}$$

The maximised likelihood ratio (MLR) for the test of $\mathcal{M}_0$ versus $\mathcal{M}_s$ can be written using the Normal approximation (Equation 19) as

$$\begin{aligned}
R_{\boldsymbol{s}} &\sim \exp\left\{\frac{n}{2}\left(\theta^{(0)}_{\mathcal{F}_s}-\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_s}\right)'\hat{\mathcal{I}}(\hat{\theta}^{(\boldsymbol{s})})_{\mathcal{F}_s,\mathcal{F}_s}\left(\theta^{(0)}_{\mathcal{F}_s}-\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_s}\right)\right\} \qquad (34) \\
&\sim \exp\left\{\frac{n}{2}\left(\theta^{(0)}_{\mathcal{F}_s}-\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_s}\right)'\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\left(\theta^{(0)}_{\mathcal{F}_s}-\hat{\theta}^{(\boldsymbol{s})}_{\mathcal{F}_s}\right)\right\}, \qquad (35)
\end{aligned}$$

the second line of approximation owing to the asymptotic equivalence of the Fisher and empirical Fisher information.

The ratio can be simplified because under the large *n* approximation, the MLEs of the parameters common to $\mathcal{M}_0$ and $\mathcal{M}_s$ are related through the expression (Cox and Hinkley 1974, p. 308).

$$\hat{\gamma}^{(0)} \sim \hat{\gamma}^{(s)} + \left[\{\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\}^{-1}\right]_{\mathcal{F}_0,\mathcal{F}_0} \left(\left[\left\{\mathcal{I}(\mathbf{0})_{\mathcal{F}_s,\mathcal{F}_s}\right\}^{-1}\right]_{\mathcal{F}_0,\mathcal{B}_s}\right)^{-1} \left(\hat{\boldsymbol{\beta}}^{(s)}_{\mathcal{B}_s} - \hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{B}_s}\right) \quad (36)$$

The simplification can be understood as taking the maximum of a slice through the likelihood surface, which is proportional to the density of $\hat{\theta}^{(S)}$. The conditional maximum is therefore computed using the formula for the conditional mean of a multivariate Normal distribution. Substituting and rearranging gives the form in Equation 33.

4.2.5 Statistical Analysis

The results of the machine learning analysis are reported in both accuracy and in relative importance of features. The accuracy is measured in AUC and the feature importance is measured in “gain”, “split” and permutation importances as well as Shapley values. All the importances are compared to the BMA results using Pearson correlation after scaling all values to between 0 and 1. All the analysis using machine learning were performed in the Python programming language.

The BMA risk factor analysis is reported as individual posterior inclusion probabilities for every risk factor, computed as described under 4.2.4.5. The directionality of effect was assessed using the mean beta across all chains. All the analysis using BMA were performed in the R programming language.

4.3 Results

The agnostic risk assessment carried out here includes 712 individual measurements from 426,893 participants. The dataset was split into cases and controls according to three criteria: (i) very severe cases of COVID-19 (871 cases), (ii) hospitalisation due to COVID-19 (2,681 cases) (iii) susceptibility to COVID-19 infection (77,867 cases) against the respective residual population. A flowchart of the dataset collection process is depicted in Figure 4-1

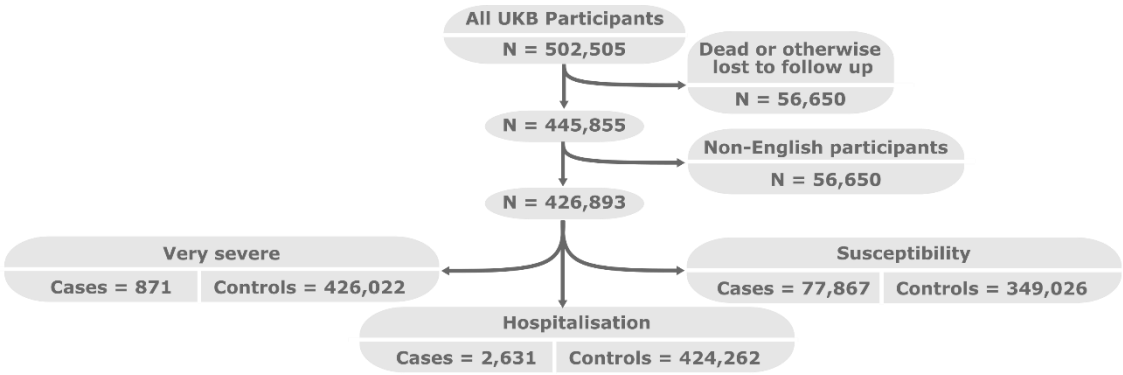


Figure 4-1 Flowchart Participants

Shows the filtering steps of UKB participants alongside the case-control-splits of the different COVID-19 outcome definitions.

Further a description of the main participant demographics through the continuous columns can be seen in Table 4-3 and binary columns are described in 4-4.

Table 4-3 Participant demographics of the dataset.

Continuous columns are shown through their mean and interquartile range. Binary columns are shows as the percentage of which the traits are present throughout the

whole dataset. The varying diseases were taken from the Charlson comorbidities indices, which groups similar ICD disease codes together.

	Mean(IQR)	Percentage present
Continuous characteristics		
Weight	77.9 (66.40-87.40)	
Height	168.5 (162-175)	
Age	70.1 (64-77)	
Waist circumference	90.1 (80-99)	
Body mass Index	27.3 (24.10-29.82)	
Binary characteristics		
Sex female		55%
Ever smoked		59%
Myocardial Infarction		3%
Congestive Heart Failure		2%
Diabetes without complications		6%
Metastatic Carcinoma		1%
Peptic Ulcer Disease		2%
Mild Liver Disease		1%
Periphral Vascular Disease		2%
Cancer		8%
Chronic Pulmonary Disease		10%
Diabetes with complications		1%
Renal Disease		2%
Paraplegia and Hemiplegia		1%
Cerebrovascular Disease		3%
Connective Tissue Disease Rheumatic Disease		2%

The same data and labels were used for phenotype prediction using machine learning in the form of a gradient boosted tree and BMA in the form of a logistic regression.

4.3.1 Machine learning

As shown in Figure 4-2, on a balanced case-control set the prediction of severe COVID-19 shows an area under the receiver operator curve (AUC) of 0.79. The prediction of hospitalisation shows an AUC of 0.75, with the susceptibility prediction being the least accurate of the three phenotypes with an AUC of 0.69. The comparison of myperformance to previous COVID-19 machine learning studies is difficult, due to differences in sample size, phenotype definition and the included independent variables. However, I can compare between the three phenotype definitions used here, where I observe a trend with more severe outcomes of COVID-19 being more accurately predicted. Phenotypes capturing more severe disease outcomes could be more biologically determined which could facilitate prediction, similar to heritability in the genome-wide association studies (GWAS) described in chapter 3. Overall, the maximum achieved AUC of 0.79 puts into question how actionable the prediction of my algorithm is for the individual patient. Possible applications could lie in augmenting clinical decision making by providing a convenient summary of health factors that contribute to COVID-19 outcome. Beyond investigating the accuracy, the predicted dependent variable shows a difference in age distribution between the varying COVID-19 outcomes, which warrants further analysis.

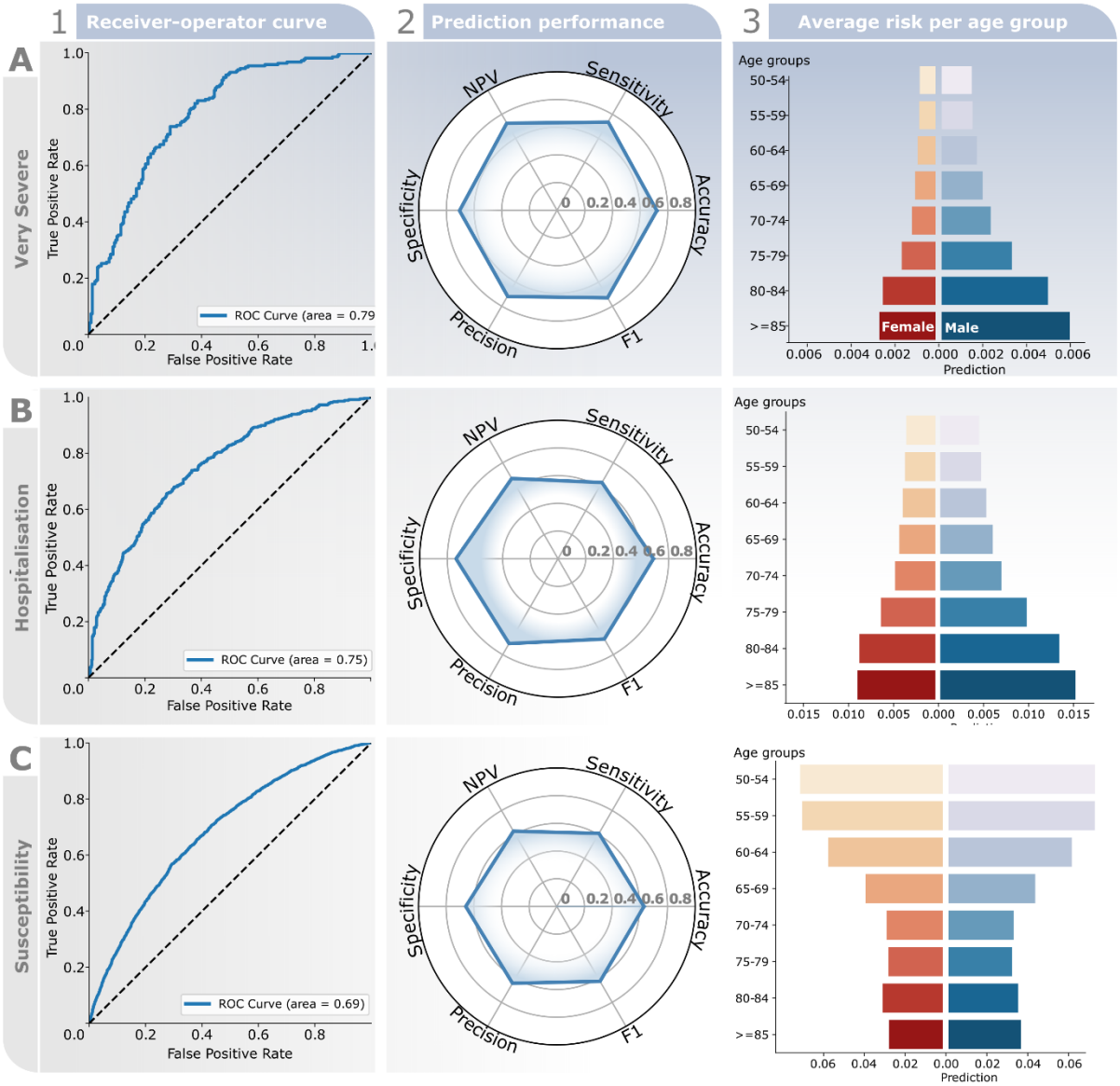


Figure 4-2 Analysis of machine learning prediction

Performance and prediction of the machine learning model for the risk of very severe COVID-19 upon infection with SARS-CoV-2 (A), hospitalisation due to COVID-19 upon infection with SARS-CoV-2 (B), and susceptibility towards COVID-19 (C). The different dependent variables are represented as rows, whereas the columns show the area under the receiver-operator-curve of the prediction (1), different performance metrics of the prediction (2) and the stratification of the prediction into age groups (3).

The machine learning prediction stratified by age shows more predicted severe cases and more predicted hospitalisations amongst older participants, which is a well-known trend that was already apparent within the first weeks of the pandemic (Bi et al. 2020; Q. Li et al. 2020; Huang et al. 2020). For susceptibility the age pyramid is inverted, with a higher prediction for younger participants to test positive for SARS-CoV-2. Older people already generally have fewer social contacts (Tran Kiem *et al.* 2021) and were specifically advised to minimise contacts during the pandemic (Cabinet Office 2020). Therefore, the lower susceptibility could be the result of lower transmission due to fewer social contacts. My findings cannot be extrapolated to those aged under 50, as there is no representation in the UKB for these age groups. Having predicted COVID-19 outcome with machine learning, I used the same data and labels for the BMA approach.

4.3.2 Bayesian Model Averaging

The BMA approach here used a logistic regression with an MCMC based sampling of risk factors to estimate a PoP that models the likely contribution of individual variables to the COVID-19 outcome, as seen in Figure 4-3.

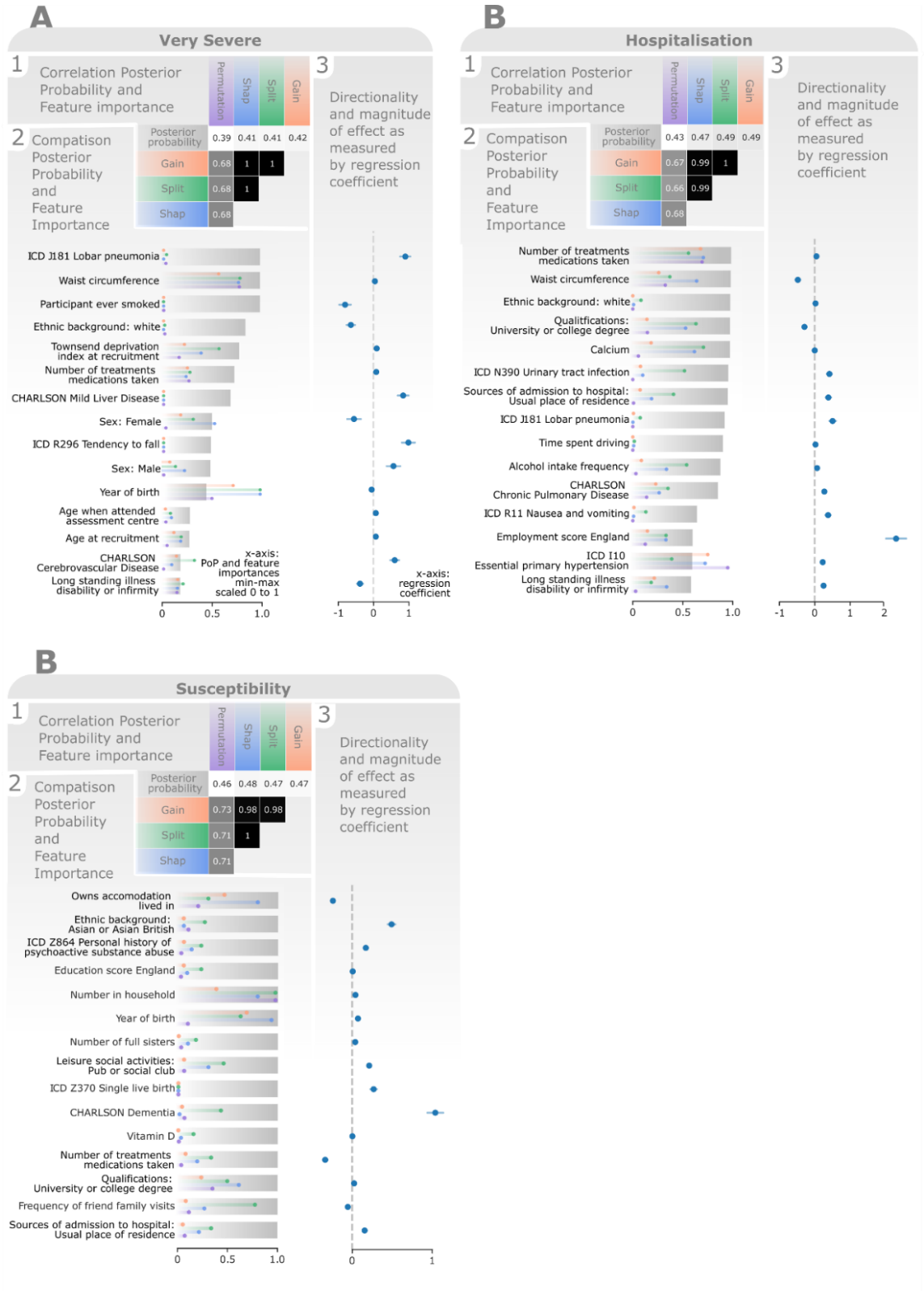


Figure 4-3 : Comparison Bayesian modelling approach and machine learning

Comparison of Posterior probability generated by the Bayesian modelling approach compared to varying machine learning feature importances using different dependent variables: severe COVID-19 upon infection with Sars-Cov-2 vs. rest of the population (A), hospitalisation due to COVID-19 upon infection with SARS-CoV-2 vs rest of the population (B) COVID-19 positive cases vs. rest of the population (C). The posterior probabilities are compared to feature importances by rank correlation (1) and by showing the posterior probabilities (grey bars) against different feature importances scaled from 0 to 1 (coloured lollipops) for the 15 independent variables with the highest posterior probability (2). The plot also shows the direction and magnitude of the effect as measured by the estimate of the regression coefficient (blue dot) and its standard error (lighter blue bar) of the logistic regression (3).

4.4 Discussion

4.4.1 Very severe COVID-19 Cases

The top scoring variable in my BMA approach to risk factor analysis for very severe COVID-19 outcomes is the ICD code J181 Lobar pneumonia caused by unspecified organism with a PoP of 0.98 and a mean beta of 0.89±0.15 (standard error). The positive beta suggests that a history of pneumonia contributes towards having severe COVID-19. The presence of past pneumonia has previously been listed as the second largest increase in hospitalisation risk among pre-existing comorbidities by Atkins *et al.* (2020). In their machine learning study Dabbah and colleagues (2021) also used Cox proportional hazard ratios and stated that pneumonia up to 12 months prior to SARS-CoV-2 infection is one of the most important predictors of mortality. Lobar

pneumonia, as opposed to bronchopneumonia and atypical pneumonia, is an infection within the lobe of a lung, where COVID-19 associated pneumonia is also located (Haseli *et al.* 2020). Normally there are distinctive immune responses to COVID-19 localised within the lungs (Szabo *et al.* 2021), which could be impaired by previous lung injury. As the UKB data only contained diagnoses made prior to the pandemic, my results indicate that a past episode of pneumonia will lead to increased risk of very severe COVID-19 even when not immediately preceding SARS-CoV-2 infection.

Whether the participant has ever smoked was the next highest PoP (0.97) and a mean beta of 0.82±0.16. The negative beta shows the protective effect of never having smoked on severe forms of COVID-19. Despite contrasting reports early in the pandemic (Clift *et al.* 2022), smoking is listed in multiple UKB risk factor studies as contributing towards, and not protecting from, severe forms of COVID-19 (Hamer *et al.* 2020; Prats-Uribe *et al.* 2021; Clift *et al.* 2022; Didikoglu *et al.* 2021; Lee *et al.* 2021; Elliott *et al.* 2021). Two studies focused on the effect of smoking specifically using UKB data. Prats-Uribe *et al.* (2021) found that that older participants had twice the risk of death of non-smokers, although there was no observed difference under the age of 69. The researchers listed social determinants of health and deprivation as influencing both smoking status and respiratory illness which could possibly confound the association. Since my approach included many measurements of health and deprivation indices, it seems unlikely that, if artefactual, smoking should have a higher PoP than any of the presumed confounders. Clift *et al.* (2022) found higher risks of hospitalisation and mortality for smokers compared to ‘never-smokers’. Using Mendelian randomisation, the study also showed that those genetically predisposed to smoking were more at risk

of infection and hospitalisation. The detrimental effect of smoking on the outcome of COVID-19 is unsurprising given the well documented association with adverse outcomes of respiratory and cardiovascular illness generally (Arcavi and Benowitz 2004), as well as bacterial and viral infection of the lung specifically (Huttunen, Heikkinen, and Syrjänen 2011). As reviewed by Han *et al.* (2019) for influenza, smoking decreases lung immunity by damaging respiratory epithelial and immune cells and thereby suppressing epithelial antiviral pathways. Smoking also facilitates cytokine release, which may be specifically detrimental for COVID-19. COVID-19 mortality has been linked to "cytokine storms", wherein an excess of proinflammatory cytokines exacerbates respiratory distress (Ragab *et al.* 2020). In addition to depressing immunity, past or current smoking has been shown to upregulate angiotensin-converting enzyme-2 (ACE2), the receptor by which SARS-CoV-2 enters human cells (Brake *et al.* 2020). Increased expression of ACE2 could lead to more opportunities for host cell entry by SARS-CoV-2. There have been many previous studies that indicate the detrimental effects of smoking on COVID-19 outcome. The strength of my study lies in the model averaging aspect, which offers a more systematic approach to variable selection. This in turn shows support for the negative effect of smoking upon SARS-CoV-2 infection.

Waist circumference (in cm) had the same PoP as never having smoked (0.974), but showed a positive beta of 0.03±0.003, indicating that a larger waist circumference predisposes patients to severe forms of COVID-19. Waist circumference in tandem with obesity is a frequently found predictor of severe forms of COVID-19 in UKB studies (Peters, MacMahon, and Woodward 2021; Freuer, Linseisen, and Meisinger 2021; Aung et al. 2020; Gao et al. 2022; J. Li et al. 2022; K. C.-Y. Wong et al. 2021; Zhu et al. 2020b;

Yates et al. 2021). Obesity is also a common comorbidity for many medical conditions and is associated with mortality from a variety of causes (Khaodhiar, McCowen, and Blackburn 1999). Similar to smoking and relevant to COVID-19, obesity increases circulation of inflammatory cytokines among other disease-contributing factors (Lockhart and O'Rahilly 2020). Interestingly however, the next highest measure of obesity was the right leg fat percentage showing a PoP of 0.02, with Body Mass Index (BMI) having a PoP of 0.02. The data presented here is indicative of visceral fat that wraps around organs, as opposed to subcutaneous that lies directly beneath the skin (Matsuzawa *et al.* 1995), being a more informative measure for severe COVID-19 than general obesity. The finding suggest that fat distribution is important, as visceral fat is found in the intra-abdominal cavity (Matsuzawa *et al.* 1995) and not, for example, around legs or arms which are measured in other obesity associated UK Biobank variables. This association has been previously investigated with UKB data by Gao and colleagues (2022), but without finding a causal relationship of central fat distribution. Freuer, Linseisen and Meisinger (2021) reported similar findings, where the impact of BMI was stronger than the amount of visceral fat in their analysis. Whereas studies based on UKB data have not found a role of visceral fat, several other studies have supported its role over BMI in COVID-19 outcome (Malavazos *et al.* 2022; Khalangot *et al.* 2022; Petersen *et al.* 2020; Bunnell *et al.* 2021; Ogata *et al.* 2021; Favre *et al.* 2021). Favre *et al.* (2021) showed that ACE2 expression in visceral fat was positively correlated with BMI, which is not true for subcutaneous fat deposits, which could suggest a mechanism leading to increased severity. Visceral fat also releases about three times more of the cytokine Interleukin-6 than subcutaneous fat (Fontana *et al.* 2007), which

is the specific cytokinin linked to severe COVID-19 (Vatansever and Becer 2020). Other than immunological consideration, more mechanical factors could play a role in COVID-19 severity. Obesity generally, and specifically fat located in the intra-abdominal cavity, leads to lower respiratory rates (Burki and Baker 1984). Visceral fat restricts movement of the diaphragm and decreases lung volume by pushing onto the lungs (He *et al.* 2021), which could exacerbate the hypoxia caused by COVID-19 (Földi *et al.* 2021). In summary, waist circumference, like smoking, appears to aggravate the adverse effects of COVID-19 in a variety of manners.

Generally, my analysis indicates that previous damage to the lungs, either by having had pneumonia or by a history of smoking, can lead to very severe COVID-19 upon infection with SARS-CoV-2. Beyond the three main factors discussed here, multiple other factors like ethnicity, social deprivation, medication, comorbidities, and sex were amongst the top factors in my analysis, which have also been commonly described in other studies (see Table 4-1). Some of these factors will be discussed in the rest of this chapter as they ranked higher in other phenotypes, whilst others will not be further discussed in the interest of brevity. Apart from investigating top factors, an approach including a comprehensive set of measurements also offers the opportunity to analyse the omission of factors within the variables with the highest PoP. Diabetes mellitus for example, which was described as an important comorbidity in early clinical studies (Huang et al. 2020; F. Zhou et al. 2020), only showed a PoP of 0.04 in my analysis. The indicated low contribution to severity suggests that diabetes itself is not fundamental in severe COVID-19. At least in the age groups investigated here, diabetes could be a confounder of risk analysis, as the disease correlates with both age and obesity (J. Luo

et al. 2020), both being independently associated with increased severity. Further, studies devoted to specific comorbidities such as periodontal disease (H. Larvin *et al.* 2021; Harriet Larvin *et al.* 2020) and mental disorders (Y. Wang et al. 2021; Kolin et al. 2020) were not supported by my analysis. The work carried out here shows that model-averaged risk factor assessments can highlight different risk factors compared to the quickly amassed wealth of COVID-19 studies published almost in real time, some of which offer diverging conclusions.

4.4.2 Hospitalisation due to COVID-19

The top scoring risk factor for hospitalisation due to COVID-19 was the number of treatments and medications taken with a PoP of 1.0. The positive mean beta of 0.05 ± 0.01, indicates that the more medications or treatments have been taken, the more likely the hospitalisation upon infection with SARS-CoV-2. The association should theoretically not be an artefact of my phenotype definitions, as only participants that were hospitalised specifically due to COVID-19 symptoms are included. Whether this was perfectly recorded in every hospital that supplied test results to UKB is unclear. The influence of medication on COVID-19 disease trajectory has been investigated using UKB data in numerous studies (Kolin et al. 2020; Ho, Petermann-Rocha, et al. 2020; McQueenie et al. 2020; K. C.-Y. Wong et al. 2021; H. Ma et al. 2021; Lehrer and Rheinstein 2021b; Xiang, Wong, and So 2021). Hypertensive medication was specifically investigated using UKB data due to the known risk factor of hypertension in COVID-19 and was found to have an association with sex-specific differences in risk (Y. Ma *et al.* 2021). One UKB study was also specifically devoted to Renin-angiotensin-aldosterone system inhibitors because of their influence on ACE2 and found to have protective

effects (Safizadeh *et al.* 2021). The top ranked variable here does not allow for such detailed insight but could rather be a measure of the general health of a participant. The more treatments or medications a participant receives, the poorer their health and the less likely they are to successfully fight off a COVID-19 infection.

The next highest scoring risk factor was, like in very severe COVID-19, waist circumference with a PoP of 0.99 and a mean beta of 0.02±0.002, underlining the adverse effects of visceral fat on COVID-19 immunity. The third ranked covariate was white ethnicity, also with a PoP of 0.99, and a mean beta of -0.49±0.07, suggesting that being white is protective against hospitalisation due to COVID-19. Ethnicity is a well-described common finding of risk factor assessments based on UKB data (Lee *et al.* 2021; Patel *et al.* 2020; Fatima *et al.* 2021; Woodward, Peters, and Harris 2021; Elliott *et al.* 2021; Kolin *et al.* 2020; S. Li *et al.* 2021; Sattar *et al.* 2020; Razieh *et al.* 2020; Lassale *et al.* 2020; Raisi-Estabragh *et al.* 2020) and has been attributed to the multiple detrimental effects of institutionalised racism which minorities are burdened with. Social deprivation is a candidate for being the underlying causal factor, but the adjustment for Townsend Deprivation Index could not fully ameliorate the adverse effects of being non-white in previous studies (Patel *et al.* 2020; Raisi-Estabragh *et al.* 2020; Lassale *et al.* 2021). My approach corroborates these results, finding that a non-white ethnicity is a stronger predictor than Townsend index (PoP=0.002). The automatic integration over all variables and exclusion of confounders allows me to readily interrogate potential correlating factors, which is a further benefit of my approach. Controlling for cardiometabolic factors, housing situation, blood biomarkers, BMI, cardiorespiratory comorbidities, and behavioural factors could not attenuate the

association with COVID-19 outcome in previous studies (Raisi-Estabragh *et al.* 2020; Patel *et al.* 2020; Lassale *et al.* 2021). My results support this for all aforementioned factors except obesity (PoP waist circumference = 0.98) and cardiorespiratory disease (Pop=0.93), as I find waist circumference and chronic pulmonary disease as additional risk factors independent from ethnicity. Differing genetic predispositions according to race have been proposed as contributing to hospitalisation but could not be verified (McCoy *et al.* 2020). The relationship could be due to higher rates of infection in minorities, which has been indicated in meta-analyses (Sze *et al.* 2020) and is supported by my analysis of COVID-19 susceptibility (see 3.1.3). The elevated transmission could stem from housing situations with more crowded living arrangements (Martin *et al.* 2020) or shared facilities (Jing *et al.* 2020). Factors associated with occupation that vary according to ethnicity have also been implicated, such as working in shifts (Maidstone *et al.* 2021), in jobs deemed essential, not being able to work remotely or being employed in crowded environments with frequent exposure during or whilst commuting to work (Sze *et al.* 2020). The debate as to how ethnicity is causal for COVID-19 hospitalisation is currently inconclusive. However, some studies suggest increased transmission could be causal and this explanation is supported by my susceptibility analysis, where housing situation is also a prominent risk factor.

Beyond the top three factors mentioned above, there are some non-intuitive risk factors amongst the highest PoP, like whether participants had college degrees (PoP=0.99), source of admission to hospital (PoP=0.96) or time spent driving (PoP=0.91). Upon closer inspection these factors have all been previously linked to either general health or COVID-19 outcome. Educational attainment is closely linked to socio-economic

standing and has been linked with COVID-19 outcome in a dedicated Mendelian randomisation study, albeit without presenting a mechanism (Yoshikawa and Asaba 2021). My study does not support socio-economic standing as an independent risk factor, with Townsend deprivation index having a PoP of 0.002. Jian and colleagues (Jian *et al.* 2021) have put forward factors correlating with education, such as overcrowding, poor housing, and hygienic practices. This is partly refuted by my work because it only has a PoP of 0.02 in my hospitalisation analysis, whereas house ownership is among the top factors for susceptibility. G. H.-Y. Li *et al.* (2021) suggest education to be linked to higher knowledge and health literacy, leading to risk aversion and stricter adherence to medical advice, which is difficult to investigate using UKB data. Time spent driving was shown to be an indicator of unhealthy lifestyles in UKB participants (Mackay *et al.* 2019), an association not supported by my analysis which assigns health score a PoP of 0.01. Time spent driving could be an indicator linked to occupation, due to commuting to work being more frequent in front line work and thus contributing to risk of hospitalisation. Being admitted to hospital from home has an associated mean beta of 0.4, which could indicate that people previously having been admitted to hospital are more likely to be hospitalised to COVID-19 as well, possibly due to generally being of ill health. COVID-19 susceptibility

In the COVID-19 susceptibility analysis the top ranked PoP was owning the apartment lived in, either directly or by someone in your household, with a PoP of 0.99. The negative mean beta of -0.24±0.02 shows that participants that own their accommodation are less susceptible to SARS-CoV-2 infection. As within-household transmission proved to be the most common mode of COVID-19 transmission (Shen *et*

al. 2020), the relation between homeownership and positive COVID-19 tests is to be expected. Assuming homeowners are more frequently the sole occupants of their home than renters, they are likely to have fewer social contacts and decreased household transmission, especially in quarantine. The importance of household transmission is also shown through the high PoP of number in household at 0.99. The effect of housing on COVID-19 has been specifically addressed by Gillies and colleagues using UKB data, albeit on severity (Gillies *et al.* 2022). Poor housing has further been linked to increased COVID-19 transmission in the US (Ahmad *et al.* 2020). House ownership is suggestive of higher income compared to renting, with higher socio-economic standing having been independently linked to decreased exposure to the virus (Beale *et al.* 2021). Differential susceptibility to COVID-19 as an effect of homeownership seems to be one of the many disadvantages suffered in the pandemic by those of lower socio-economic standing. The connection between deprivation and COVID-19 susceptibility is apparent when examining the other top ranked variables. These are ethnicity (PoP = 0.99), history of psychoactive abuse (PoP = 0.99), education score (PoP = 0.99), number in household (PoP= 0.99), number of full sisters (PoP = 0.99) and university and college degrees (PoP = 0.0.96), which can all be described as capturing or correlating with socio-economic standing.

Being of Asian or Asian British descent is the second highest PoP with 0.99 and a mean beta of 0.49±0.05, making individuals of this ethnic background more likely to test positive for COVID-19 compared to other ethnicities. Again, socio-economic factors could be causal, however the specific ethnicity observed here differs with the more general non-white association in very severe COVID-19. The specifically higher risk of

South Asian heritage compared to all other ethnicities was previously shown amid the second wave of the pandemic in the UK (Mathur *et al.* 2021). The study attributes household size and multi-generational living in South Asian communities to the higher transmissibility, which is supported by other factors uncovered, like homeownership and number in household.

Between all three phenotypes that measure the impact of COVID-19 on my society, socio-economic standing and ethnicity seem to be a common factor alongside medical markers of ill health. The usefulness of my approach is not limited to the analysis of individual risk factors, but also allows for the comparison to off-the-shelf feature importance measures coming from machine learning.

4.4.3 Comparison Feature Importance and Posterior Probability

3 shows the different machine learning feature importances alongside the posterior probabilities created by BMA. Across all phenotypes and all feature importances, there is a general trend wherein non-binary variables like waist circumference or number of medications are associated with higher feature importance. In comparison, binary variables like ICD codes and dummy variables like ethnic background rank lower in relative feature importance. The difference becomes more apparent when comparing means across all phenotypes and variables. The means of all posterior probabilities between binary and continuous variables differ by 9.38% (binary = 0.478, continuous = 0.437), whereas the feature importance means differ by 198.07% (binary = 0.104, continuous 0.206). It is a common occurrence that continuous features are ranked higher in importance of decision machine learners like gradient boosted trees. Decision tree based learners exhaustively explore splits over all variables within a dataset (Loh

and Shih 1997). Therefore, there is a higher probability of finding a better split by chance amongst the many possible splits in a continuous variable when compared to a binary variable, which only offers one possible split (Hothorn, Hornik, and Zeileis 2006). When using feature importance as a proxy for likely influence of variables on COVID-19 outcome, researchers must keep in mind that binary variables are biased against from the outset. This becomes particularly problematic when investigating comorbidities, as those are usually defined by being either present or absent in the patient. Bias against binary variables, however, is not the only concern.

Further doubt is cast on the utility of relative feature importance for risk factor importances by comparing rank correlations across different phenotypes. Whereas the feature importance measures of gain, split and Shapley values show good correlations with each other across all phenotypes (range = 0.98-1.0), they do not correlate well with PoP (range = 0.41-0.49). Permutation importance seems to be an outlier of both machine learning feature importances (range = 0.66-0.71) and PoP (range = 0.39-0.46). My comparison of relative feature importance with BMA suggests that interpreting high ranking features as risk factors, as has been done in previous UKB studies (Dabbah et al. 2021; K. C.-Y. Wong et al. 2021), is questionable. My comparison supports the assertion that without further efforts to disentangle chains of correlation, the utility of machine learning lies in classification and not inference.Beyond the comparison to machine learning, the comparison of this method to other risk factor prediction method like logistic regresson, poisson regression and cox models would have been helpful for outlining the differences between established methods and my method. Due to the expected great amount of computational and time requirements of implementing

classical models on all biomarkers in UKB, I was not able to deploy these methods because of time constraints.

4.5 Conclusions and Outlook

The difference between the top-ranking variables in feature importance and BMA risk analysis underlines the need for a dedicated risk factor analysis. Common to all my risk analyses are descriptors of socio-economic standing. Factors describing educational attainment, living situation, ethnicity, and deprivation ranking high in all analyses underlines how the pandemic trajectory traces existing lines of the societal wealth divide. Not all socio-economic factors contribute to all phenotypes, as the susceptibility analysis uniquely emphasises the involvement of accommodation, possibly by influencing household transmission. The analysis of very severe COVID-19 and hospitalisation generally unveils risk factors describing ill health, specifically obesity, which do not feature in the susceptibility analysis. Alongside measures of obesity, variables indicating some form of lung injury are present in both analyses of COVID-19 severity. Very severe COVID-19 shows a history of pneumonia and previous smoking among the top-ranking factors and the Charlson code for chronic pulmonary disease and pneumonia are among the highest-ranking variables in hospitalisation. Intuitively, a lung that has previously been damaged is more likely to decline when attacked by a Sars-Cov-2 infection than a healthy lung. Applying standardised methodology with differing phenotypes shows how the BMA can be deployed irrespective of the dependent variable, however the utility of the approach lies beyond comparisons within my study but rather by contrasting it with other research.

I argue that my agnostic approach ameliorates concerns about confounders and selection bias through the inclusion of a broad set of variables, the validity of which can be shown empirically. Well-described risk factors for which there is a consensus in the literature to be detrimental for COVID-19 outcome, like obesity (12 UKB studies, see Table 4-1), socio-economic standing (10 UKB studies), ethnicity (9 UKB studies) and smoking (5 UKB studies) are re-discovered in my analysis. Supporting an existing consensus in literature lends credibility to the adequacy of my approach for a bias-free risk assessment. This has encouraged me to investigate the diverging hypotheses suggested by a flood of studies brought about by an urgent demand for insight into COVID-19.

My agnostic and comprehensive risk assessment for COVID-19 outcome using UKB data supports some of the previous UKB studies, while refuting others. Comorbidities that were previously outlined to be contributing to a COVID-19 associated decline of health, like diabetes mellitus (Atkins et al. 2020; F. Zhou et al. 2020; K. C.-Y. Wong et al. 2021; Elliott et al. 2021; Gao et al. 2022), periodontal disease (Harriet Larvin *et al.* 2020; H. Larvin *et al.* 2021), mental disorders (Y. Wang et al. 2021; Kolin et al. 2020) and asthma (Zhu *et al.* 2020a; Lodge *et al.* 2021) are not supported by my analysis. Factors with only a few supporting studies using UKB data, like air pollution contributing to COVID-19 mortality and susceptibility (Travaglio *et al.* 2021), or glasses reducing COVID-19 susceptibility (Lehrer and Rheinstein 2021a) are also not supported by my analysis. The position taken only by H. Ma *et al.* (2021) among all UKB studies, that habitual Vitamin D use is protective of infection is corroborated by my analysis, which has the blood biomarker for Vitamin D ranked 11th (PoP=0.98) in susceptibility. This finding goes

against other UKB studies refuting the work of Ma and colleagues (C. E. Hastie, Pell, and Sattar 2021; S. Li et al. 2021; Elliott et al. 2021). I additionally describe novel factors not previously observed in UKB studies, such as the effect of a history of pneumonia on developing very severe COVID-19 (1st rank, PoP = 0.975) and the effect of blood calcium on COVID-19 hospitalisation (5th rank, PoP = 0.984). The effect of prior pneumonia (Mouliou, Kotsiou, and Gourgoulianis 2021) and calcium on COVID-19 severity (Alemzadeh *et al.* 2021) is supported by research based on other data sources. Similarly, I find waist circumference the superior descriptor of COVID-19 compared to other measures, contrary to other UKB studies (Gao *et al.* 2022; Freuer, Linseisen, and Meisinger 2021), but in line with studies based on other data (Malavazos *et al.* 2022; Khalangot *et al.* 2022; Petersen *et al.* 2020; Bunnell *et al.* 2021; Ogata *et al.* 2021; Favre *et al.* 2021). BMA allows for both the discovery of new risk factors and supporting or refuting existing risk factor studies without being subject to the same concerns, which makes it widely applicable in epidemiological settings.

Despite its demonstrated utility here and in other recent studies with different foci (Mu, See, and Edwards 2019; Yimer *et al.* 2021; Hinne *et al.* 2020), BMA has not been used for COVID-19 analysis in the UKB prior to this study. Neither is it a commonly deployed tool in risk factor analysis of disease generally (Mu, See, and Edwards 2019), which is a fatal omission in the epidemiological toolkit as variable inclusion confounds conclusions (Kraemer *et al.* 2001). *Ad hoc* manual variable selection based on prior experience is difficult to audit and can lead to irreproducible results (Sauerbrei *et al.* 2020). In the currently predominant analysis methodology, manual or other variable selection is followed by association tests, which can be reductionist in their over-reliance on the p-value given that it offers limited information (Halsey 2019). My approach addresses

both criticisms, as BMA produces readily interpretable posterior probabilities, and the agnostic approach mitigates concerns about the inscrutability of variable selection. My agnosticism is expressed in the definition of the prior, where I assume all independent variables equally likely to be associated with the dependent variable, which automates the control for confounders. The utility of my methodology is not limited to COVID-19, as demonstrated here by using 3 different phenotypes, and can easily be reused for any disease or binary trait of interest. To facilitate the use of my method for different research questions, I have produced a computationally efficient approach by exploiting the Normal approximation to the likelihood that is justified by the biobank-scale data. Beside facilitating feasibility, it also opens the door for PheWAS-type approaches for non-genetic risk factors to augment GWAS approaches. Similarly choosing the g-prior reduces the required computational resources as it depends on only two hyperparameters, $\mathcal{M}$ and h. First principles indicate that h should be constant as the data grows large, and therefore negligible compared to n, which is convenient for big data approaches. Sensitivity analysis preceding the final risk factor analysis confirms this by showing the results were robust to h=1 vs h=0.1 (data not shown). A further departure from standard practice was the inclusion of factors by encoding every level separately. The collinearity of the resulting dummy variables does not trip up the resulting risk analysis in BMA, which is a benefit over other methods as it allows for a more fine-grained analysis. For example, I was able to delineate non-whiteness as a relevant factor instead of being reduced to saying ethnicity contributes to risk generally. Furthermore, since each level is associated with a single degree of freedom, the criteria for model inclusion are more easily met than when imposing the constraint that all levels of the factor must be included or excluded simultaneously. Preliminary results confirmed the utility of this approach as without encoding levels separately, multi-level factors were rarely among the highest scoring risk factors. By

designing tools with computational efficiency in mind, I hope to make my BMA method more accessible.

In this study I have laid out the benefits of BMA over other methods of risk factor discovery on first principles, which I subsequently demonstrated empirically in a COVID-19 framework. My computationally efficient method, unconstrained by concerns for confirmation bias and confounders, provides a rigorous and standardised tool to risk factor discovery usable for any disease of interest.

Chapter V

Discussion

Overview

The big data era of infectious disease research has already been marked by great challenges such as the coronavirus disease (COVID-19) pandemic and the rise of antimicrobial resistance, which have been fought on many fronts using techniques from statistics and machine learning. In this closing chapter I argue how I have contributed to big data analysis of infection by motivating, designing, and applying inference and prediction tools for generating insight into COVID-19 and Campylobacteriosis. I briefly summarise the merits of my *C. jejuni* source prediction, the insight generated through genome-wide association studies (GWAS) based on source prediction, and the comprehensive COVID-19 risk assessment generated by combining prediction and inference. I further highlight the limitations of my tools and the findings they generate whilst proposing means to overcome them through additional *in silico* analysis and by experimental validation of the generated hypotheses. In line with my aim to not only generate domain knowledge but also provide tools that can be repurposed to novel infectious disease settings, I suggest future areas of application for my methodology. The integration of my algorithms in public databases has the potential to make my work accessible for researchers, including those that generated the data which enabled my analysis. My closing arguments return to the topics covered in introductory chapter and consider how bridging the divide between the principles of algorithmic and data modelling can unlock the potential of big data in addressing current and future challenges in infectious disease research.

5.1 Summary of the Thesis

Alongside enormous infectious disease challenges, the 21st century has also witnessed the advent of big data, with the increasing variety, volume, and velocity of information offering new lines of investigation to combat novel and ancient pathogens. Various methods have been designed to tap into the potential of big data with two largely separate goals and tools to achieve them: statistical inference and machine learning prediction. My aims were to motivate, develop and apply the right tools for the right job, which can also be easily re-purposed to generalise across other infections. By outlining the theoretical background of the methodology alongside their goals and limitations in the introductory chapter, I established the tool choice of the subsequent chapters. In Chapters 2, 3 and 4 I used these tools to contribute novel insights into campylobacteriosis and coronavirus disease (COVID-19):

- In **Chapter 2** I used a gradient boosted classifier to predict the source of Campylobacteriosis from publicly available sequences deposited in PubMLST. The resulting machine learner, aiSource, improved the accuracy of the previously most frequently applied method by 33%. Beyond improving accuracy, aiSource widened the input spectrum of source attribution from multi-locus sequence typing (MLST) to core-genome MLST (cgMLST) and whole genomic sequences (WGS), allowing individualised source attribution of genetically unique isolates for the first time. Investigating predicted sources within the generalist lineage ST-21 showed varying host affinity within closely related sublineages, which raised the question of whether genetic changes contribute to a switch in host affinity.

- Building on the fine-grained prediction of aiSource, **Chapter 3** sought genetic variation in *C. jejuni* underlying host affinity using WGS from PubMLST. I found the host specificity between ruminants and chickens associated with genetic variation observed in polyphosphate uptake and use. A commonality of the adaption of both chicken and ruminant isolates to the human microenvironment was the discovery of genes contributing to an increased ability to survive meat processing and entering of host endothelial cells. These two abilities also mediate pre-adaption to transmission to humans through a fluoroquinolone (FQ) resistance conferring mutation in the *gyrA* gene of chicken strains. As I only discovered this association in strains transmitting from chicken to humans, the *gyrA* mutation could be an underlying factor of chicken isolates being the only niche specialists that readily transmit to humans as we have observed in Chapter 2.
- The combination of prediction and inference was the basis for **Chapter 4**, where I used gradient boosted trees and Bayesian model averaging (BMA) to predict risk and uncover risk factors for COVID-19 outcome and susceptibility from UK Biobank (UKB) data. Living situations influencing household transmission were the most important risk factors for susceptibility. Known factors of COVID-19 disease severity, like ethnicity, socio-economic standing, obesity, and markers of ill health were rediscovered alongside less well-examined factors like history of pneumonia. By using an agnostic approach, I was able to evaluate findings of previous UKB analyses. My results supported some previous findings, like the

importance of visceral fat in the severity of disease, and challenged others, like the influence of air pollution on COVID-19 mortality risk.

The interpretation of my findings and their contribution to their respective research domains have been laid out in the result chapters. I therefore have only briefly summarised my findings and will subsequently focus on the limitations of my work.

5.2 Limitations

There are several limitations to be highlighted throughout the result chapters of this thesis:

- In **Chapter 2** rapid host switching makes source attribution particularly challenging, because it generates overlapping gene pools, such as those between the cattle (85% of correctly identified source labels (accuracy), see Figure 2-4) and sheep niches (57% accuracy). Errors caused by the underlying population biology are hard to avoid, but my approach also suffers from the data-greedy nature of training, as is generally the case in machine learning. Whereas well sampled classes like chicken can accurately be identified (93% accuracy), the more sparsely sampled environmental niche is more difficult to attribute (77% accuracy). With the inclusion of more data, especially of the less well sampled classes, the predictive performance should improve.

  Other than improving accuracy, the processing of inputs could be revised. In widening the input spectrum of source attribution my approach subsequently focused on cgMLST. The allelic profile must be assigned from raw WGS prior to the analysis and cannot leverage information in the accessory genome, where host specific genetic factors have been shown to lie (Sheppard

et al. 2013). My WGS approach capturing information beyond the core genome could not leverage the full genomic information: The WGS were k-merised and dimensionality reduction, in the form of feature selection, was used to accommodate the finite computational resources. A more effective capture of genomic data or more bespoke algorithm design which uses the genotype information better would have great potential for source attribution as WGS becomes increasingly available through ever-declining sequencing costs.

My tool choice is also not without limitations. Machine learning can draw criticism for improving prediction accuracy without offering more biological insight. The use of aiSource by itself without reaping the benefits of individual attribution offers few advantages over existing methods beyond increased accuracy. I tried to address this by investigating the genetic basis of host affinity switching by using the labels as a starting point for Chapter 3.

- **Chapter 3** uses the prediction of aiSource and thus alleviates the criticism that limited biological insight can be gained from machine learning. However, as a tool for statistical inference aiSource is subject to its own limitations. To control for possible confounders in sequencing set-up and assembly, I was required to limit the dataset to three bioprojects with a similar study design. The filtering halved the number of samples used in Chapter 2 and limited the analysis to just samples collected in the UK. Without further analysis, this filtering restricts the conclusions about the mechanisms of C. jejuni adaptation to isolates from the UK. This limits the interpretation of my hypothesis that FQ resistance pre-adapts C. jejuni to humans, where US samples could have allowed verification through

investigating the effects of the American FQ ban. Further experimental work could also investigate whether FQ resistance, increased survival through poultry processing, and increased epithelial cell invasion is mediated by increased biofilm formation triggered by the *gyrA* mutation. Other resistance conferring mutations for FQ and other antimicrobials would also have to be screened for fitness, to confirm the of the pre-adaption of only FQ resistance among all antimicrobials and verify that my findings are not sampling artefacts. If the finding that FQ resistance increases chicken to human isolate transmission is confirmed experimentally, suggesting a shift to different antibiotic use for chicken processing. Convincing the poultry production industry to cease all antimicrobial use would be more difficult, as the dangers of pervasive antibiotic use in increasing levels of resistance are already known and have not managed to do so (Torres *et al.* 2021).

Beyond FQ resistance, all mechanisms of adaption I put forward require further experimental validation. The involvement of the polyphosphate pathway, for example, is only indirectly suggested through interaction with other genes, requiring further proof. Experimental validation would conclude the knowledge discovery process started by prediction in Chapter 2 to inference in Chapter 3.

- The dual approach used in **Chapter 4** has similar shortcomings to those listed above. The machine learning algorithm itself offers limited insight about COVID-19 risk, and the comparably low accuracy makes use in medical settings questionable.

In the BMA approach, rigorous data cleaning was required in the form of exclusion of columns with a high degree of missing data, or with too many levels and values had to be imputed and repeated measurements combined. Although I have designed a method that can be readily re-applied as is, the data cleaning must be applied according to the limitations of the data at hand. A re-application would additionally require an adjustment of the priors, for example μ which defines the expected number of associated risk factors.

The re-application of the BMA implementation to other research questions is currently limited to binary dependent variables, which is sufficient for disease states, but could be amended for continuous variables which would broaden the spectrum of applications. This would require modestly generalising the existing code. Although the computational feasibility of the approach was helped by a Normal approximation to the likelihood and the use of a conjugate g-prior, which requires only two hyper-parameters, the required computational resources are still considerable. The risk factor analysis for the UKB dataset consisting of 426,893 rows and 1,143 columns, using 50 chains and 100,000 iterations, took three days of computation on 50 CPUs at 2.4 GHz each. The length of the Markov chain Monte Carlo (MCMC) chain is the time-limiting factor and further consideration for better mixing leading to shorter chains could lower the computational requirements. Borrowing concepts of machine learning could be helpful, as an early stopping mechanism based on measuring the mixing process could shorten computational time as it does when minimising predictive loss. The output of the BMA analysis, which is currently given only as a PoP,

might benefit from the definition of a significance cut-off to automate the selection of interesting factors, which would further reduce the subjectivity of the analysis.

The risk factors suggested by the BMA analysis would also have to be confirmed through bespoke experimental design for individual factors, especially for patients under 50 as all participants in UKB were over 50 at the time of analysis. Considering the discovered risk factors capture ill health and socio-economic standing, verification would be limited to observational instead of the interventionist experiments proposed earlier in the *Campylobacter* context.

The need for lab work downstream of inference methods like genome-wide association studies (GWAS) and BMA are not shortcomings of the tool design, but inevitabilities of inferring knowledge about an unseen process. Discovering causal relationships may lie at the heart of statistical inference, but causality cannot be outright demonstrated using observational data (Pearl 2009). Designing experiments that investigate how changing one independent variable affects a dependent variable can deliver proof of causality. Experimental work, albeit sometimes unjustly overlooked in the analysis of large scale computational work, is closely integrated in the cycle of big data analysis as is illustrated by my campylobacteriosis analysis: lab generated sequences are used in source prediction which provides starting points for inference that generates hypotheses which are confirmed by lab research. Beyond elaborating on how computational work like mine can benefit from *in vitro* studies, I also want to underline how my research could benefit and feed back into experimental research.

5.3 Outlook and Conclusions

Irrespective of the aims of prediction versus inference, my work is indebted to the maintainers of large, well curated, databases like PubMLST and UKB, and the researchers that contribute data to them. With few adjustments my big data analysis methods could be automated and redeployed within public databases, thus lowering the threshold of statistical expertise and programming knowledge required for their use. The speed of prediction and low computational requirements for the algorithm developed in Chapter 2 favour its integration into PubMLST. This would allow microbiologists and clinicians to automatically assign sources to human *C. jejuni* and *C. coli* isolates upon upload. AiSource, available from https://github.com/narning1992/aiSource, has the built-in capability to be retrained given any label, allowing the user to predict different phenotypes like country of origin or year of sampling for *Campylobacter* or any other microbial cgMLST sample. As was demonstrated in Chapter 2, the predicted labels can be used as phenotypes for microbial GWAS. The output of the algorithm trained on cgMLST data also provides a convenient summary of the genetic background and could be used as a covariate in GWAS for an enhanced control of population stratification. Similarly, my agnostic BMA approach can be repurposed for conducting non-genetic risk analysis using different disease labels. UKB is an ideal starting point as a comprehensive health dataset on which I have demonstrated the feasibility of my approach and could make pre-processed data available. The output of the logistic regression could further be used as a covariate to control non-genetic confounders in GWAS, which would complement the use of the aiSource prediction as

a covariate. Both approaches together could help to enhance the utility of databases for GWAS, particularly as efforts linking microbial isolates and sequences to UKB participants are underway (Armstrong *et al.* 2020; Hilton *et al.* 2020). GWAS of infectious disease could be applied to understand the impact of microbial and human genetic factors in unison, where BMA can control non-genetic human confounders of disease and aiSource population stratification of the pathogen using UKB. The accumulation of easily accessible data in large-scale databases is a major driver of progress in big data research and I hope to make my contribution by making large scale analysis more accessible. My work should not only encourage collaboration between computational and experimental researchers, but also between the communities pursuing statistical inference and machine learning.

By motivating and demonstrating the appropriate use of computational methodology in infectious disease research I show the respective strengths and limitations of inference and prediction. In Chapter 4 I specifically demonstrate that feature importance without further analysis cannot substitute statistical inference for risk factor analysis. However, I also exemplify how prediction and inference can be complementary despite differences in underlying philosophy. In Chapter 3 I show how predicting sources can give starting points of inference through GWAS and in Chapter 4 risk factor analysis is augmented by prediction of outcome to offer a complete view on COVID-19 risk. My results show how both research communities can benefit from adopting approaches of the other as is also seen in Chapter 4. The agnostic approach of including a vast array of available UKB biomarkers in the BMA approach is analogous to the data-greedy approach taken in machine learning and allowed me to investigate previous UKB

studies. Bridging the divide between the two approaches of algorithmic and data modelling could be crucial in addressing the today's most urgent research questions.

As outlined in the introductory chapter, the 21st century has brought massive challenges for infectious disease research in the rise of novel agents like COVID-19, continuing threats like malaria and HIV and the re-emergence of old diseases through the rise of antimicrobial resistance. Whilst the prevailing view towards infectious diseases in the preceding century might have been optimistic, the outlook has deteriorated since, and especially during the ongoing COVID-19 pandemic. Fortunately however, this century was also witness to the big data revolution, which can address these challenges if the potential of the ever-increasing wealth of information is tapped into. In my work, we have demonstrated this by contributing to COVID-19 and Campylobacteriosis research and laid the groundwork for making such large-scale analyses more accessible. Revisiting the questions big data address listed in the introduction – What is going to happen? What will happen if some variable is changed? What is the pattern of the underlying data? (Iwashyna and Liu 2014) – I want to restate these questions with respect to my work:

- **How is the observed data generated?** The domain of statistical inference aiming to understand data generating processes through the language of probability.
- **What data will be generated?** Which can be addressed with machine learning prediction mirroring the hidden structure of the data in algorithms.

When investigating the notion of extrapolating from the clinical past to the future from evidence based medicine in the introductory chapter (Ehrenstein *et al.* 2017), I have focused on how the evidence has changed in variety, velocity and volume, thereby

powering knowledge generation. However, using the past to glimpse into the future also offers a way forward for big data analysis in the form of two more questions:

- How can the understanding of a data generation process contribute to our imitation of it?
- How can the imitation of a data generating mechanism further our understanding of it?

I have offered a glimpse of how to answer the latter by contributing a small step along this path with encouraging results. Similar research will hopefully encourage the collaboration between both modelling communities to better address both questions. The synthesis of statistical inference and machine learning prediction might be able to unlock the enormous potential of big data, which would allow us to view the future of infectious diseases with optimism once more.

References

Abu Alfeilat, Haneen Arafat, Ahmad B.A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S. Eyal Salman, and V.B. Surya Prasath. 2019. 'Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review'. *Big Data* 7 (4): 221–48. https://doi.org/10.1089/big.2018.0175.

Affonso, Carlos, André Luis Debiaso Rossi, Fábio Henrique Antunes Vieira, and André Carlos Ponce de Leon Ferreira de Carvalho. 2017. 'Deep Learning for Biological Image Classification'. *Expert Systems with Applications* 85 (November): 114–22. https://doi.org/10.1016/j.eswa.2017.05.039.

Ahmad, Khansa, Sebhat Erqou, Nishant Shah, Umair Nazir, Alan R. Morrison, Gaurav Choudhary, and Wen-Chih Wu. 2020. 'Association of Poor Housing Conditions with COVID-19 Incidence and Mortality across US Counties'. *PLOS ONE* 15 (11): e0241327. https://doi.org/10.1371/journal.pone.0241327.

Alaa, Ahmed M., Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, and Mihaela van der Schaar. 2019. 'Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423,604 UK Biobank Participants'. *PLOS ONE* 14 (5): e0213653. https://doi.org/10.1371/journal.pone.0213653.

Aldred, KatieJ., Robert J. Kerns, and Neil Osheroff. 2014. 'Mechanism of Quinolone Action and Resistance'. *Biochemistry* 53 (10): 1565–74. https://doi.org/10.1021/bi5000564.

Alemzadeh, Effat, Esmat Alemzadeh, Masood Ziaee, Ali Abedi, and Hamid Salehiniya. 2021. 'The Effect of Low Serum Calcium Level on the Severity and Mortality of Covid Patients: A Systematic Review and Meta-Analysis'. *Immunity, Inflammation and Disease* 9 (4): 1219–28. https://doi.org/10.1002/iid3.528.

Allen, Michael Patrick, ed. 1997. 'Testing Hypotheses in Nested Regression Models'. In *Understanding Regression Analysis*, 113–17. Boston, MA: Springer US. https://doi.org/10.1007/978-0-585-25657-3_24.

Altekruse, Sean F., Norman J. Stern, Patricia I. Fields, and David L. Swerdlow. 1999. 'Campylobacter Jejuni—An Emerging Foodborne Pathogen'. *Emerging Infectious Diseases* 5 (1): 28–35. https://doi.org/10.3201/eid0501.990104.

Altman, Douglas G, and J Martin Bland. 2005. 'Standard Deviations and Standard Errors'. *BMJ : British Medical Journal* 331 (7521): 903. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1255808/.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. 'Basic Local Alignment Search Tool'. *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Amin, Hasnat A, and Fotios Drenos. 2021. 'No Evidence That Vitamin D Is Able to Prevent or Affect the Severity of COVID-19 in Individuals with European Ancestry: A Mendelian Randomisation Study of Open Data'. *BMJ Nutrition, Prevention & Health* 4 (1): 42–48. https://doi.org/10.1136/bmjnph-2020-000151.

Anahtar, Melis N., Jason H. Yang, and Sanjat Kanjilal. 2021. 'Applications of Machine Learning to the Problem of Antimicrobial Resistance: An Emerging Model for Translational Research'. *Journal of Clinical Microbiology* 59 (7): e01260-20. https://doi.org/10.1128/JCM.01260-20.

Anderson, Jana J., Frederick K. Ho, Claire L. Niedzwiedz, Srinivasa Vittal Katikireddi, Carlos Celis-Morales, Stamatina Iliodromiti, Paul Welsh, et al. 2021. 'Remote History of VTE Is Associated with Severe COVID-19 in Middle and Older Age: UK Biobank Cohort Study'. *Journal of Thrombosis and Haemostasis* 19 (10): 2533–38. https://doi.org/10.1111/jth.15452.

Andreu-Perez, Javier, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang. 2015. 'Big Data for Health'. *IEEE Journal of Biomedical and Health Informatics* 19 (4): 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362.

Ansari, M. Azim, and Xavier Didelot. 2016. 'Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree'. *Genetics* 204 (1): 89–98. https://doi.org/10.1534/genetics.116.190496.

Appasani, Krishnarao, ed. 2016. *Genome-Wide Association Studies: From Polymorphism to Personalized Medicine*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107337459.

Arcavi, Lidia, and Neal L. Benowitz. 2004. 'Cigarette Smoking and Infection'. *Archives of Internal Medicine* 164 (20): 2206–16. https://doi.org/10.1001/archinte.164.20.2206.

Argimón, Silvia, Khalil Abudahab, Richard J. E. Goater, Artemij Fedosejev, Jyothish Bhai, Corinna Glasner, Edward J. Feil, et al. 2016a. 'Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography'. *Microbial Genomics,* 2 (11): e000093. https://doi.org/10.1099/mgen.0.000093.

———. 2016b. 'Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography'. *Microbial Genomics* 2 (11): e000093. https://doi.org/10.1099/mgen.0.000093.

Armstrong, Jacob, Justine K. Rudkin, Naomi Allen, Derrick W. Crook, Daniel J. Wilson, David H. Wyllie, and Anne Marie O'Connell. 2020. 'Dynamic Linkage of COVID-19 Test Results between Public Health England's Second Generation Surveillance System and UK Biobank'. *Microbial Genomics* 6 (7). https://doi.org/10.1099/mgen.0.000397.

Arning, Nicolas, Samuel K. Sheppard, Sion Bayliss, David A. Clifton, and Daniel J. Wilson. 2021. 'Machine Learning to Predict the Source of Campylobacteriosis Using Whole Genome Data'. *PLOS Genetics* 17 (10): e1009436. https://doi.org/10.1371/journal.pgen.1009436.

Atkins, Janice L., Jane A. H. Masoli, Joao Delgado, Luke C. Pilling, Chia-Ling Kuo, George A. Kuchel, and David Melzer. 2020. 'Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort'. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 75 (11): 2224–30. https://doi.org/10.1093/gerona/glaa183.

Aung, Nay, Mohammed Y. Khanji, Patricia B. Munroe, and Steffen E. Petersen. 2020. 'Causal Inference for Genetic Obesity, Cardiometabolic Profile and COVID-19 Susceptibility: A Mendelian Randomization Study'. *Frontiers in Genetics* 11. https://www.frontiersin.org/article/10.3389/fgene.2020.586308.

Austerlitz, Frederic, Olivier David, Brigitte Schaeffer, Kevin Bleakley, Madalina Olteanu, Raphael Leblois, Michel Veuille, and Catherine Laredo. 2009. 'DNA Barcode Analysis: A Comparison of Phylogenetic and Statistical Classification Methods'. *BMC Bioinformatics* 10 (14): S10. https://doi.org/10.1186/1471-2105-10-S14-S10.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. 'A Global Reference for Human Genetic Variation'. *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Baig, Sabeeh A. 2020. 'Bayesian Inference: An Introduction to Hypothesis Testing Using Bayes Factors'. *Nicotine & Tobacco Research* 22 (7): 1244–46. https://doi.org/10.1093/ntr/ntz207.

Banerjee, Amitav, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar, and S. Chaudhury. 2009. 'Hypothesis Testing, Type I and Type II Errors'. *Industrial Psychiatry Journal* 18 (2): 127–31. https://doi.org/10.4103/0972-6748.62274.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing'. *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bansal, Shweta, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, and Cécile Viboud. 2016. 'Big Data for Infectious Disease Surveillance and Modeling'. *The Journal of Infectious Diseases* 214 (suppl_4): S375–79. https://doi.org/10.1093/infdis/jiw400.

Bashir, Daniel, George D. Montañez, Sonia Sehra, Pedro Sandoval Segura, and Julius Lauw. 2020. 'An Information-Theoretic Perspective on Overfitting and Underfitting'. In *AI 2020: Advances in Artificial Intelligence*, edited by Marcus Gallagher, Nour Moustafa,

and Erandi Lakshika, 347–58. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-64984-5_27.

Beale, Sarah, Isobel Braithwaite, Annalan MD Navaratnam, Pia Hardelid, Alison Rodger, Anna Aryee, Thomas E. Byrne, et al. 2021. 'Deprivation and Exposure to Public Activities during the COVID-19 Pandemic in England and Wales'. *J Epidemiol Community Health*, October. https://doi.org/10.1136/jech-2021-217076.

Beck, James Vere, and Kenneth J. Arnold. 1977. *Parameter Estimation in Engineering and Science*. James Beck.

Bellman, R. 1966. 'Dynamic Programming'. *Science (New York, N.Y.)* 153 (3731): 34–37. https://doi.org/10.1126/science.153.3731.34.

Bencharit, Sira, and Mandy J. Ward. 2005. 'Chemotactic Responses to Metals and Anaerobic Electron Acceptors in Shewanella Oneidensis MR-1'. *Journal of Bacteriology*, July. https://doi.org/10.1128/JB.187.14.5049-5053.2005.

Ben-Hur, Asa, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. 2008. 'Support Vector Machines and Kernels for Computational Biology'. *PLOS Computational Biology* 4 (10): e1000173. https://doi.org/10.1371/journal.pcbi.1000173.

Bi, Qifang, Yongsheng Wu, Shujiang Mei, Chenfei Ye, Xuan Zou, Zhen Zhang, Xiaojian Liu, et al. 2020. 'Epidemiology and Transmission of COVID-19 in 391 Cases and 1286 of Their Close Contacts in Shenzhen, China: A Retrospective Cohort Study'. *The Lancet Infectious Diseases* 20 (8): 911–19. https://doi.org/10.1016/S1473-3099(20)30287-5.

Birkes, D. 2005. 'Likelihood Ratio Tests'. In . https://doi.org/10.1002/0470011815.b2a15074.

Bloom, David E., and Daniel Cadarette. 2019. 'Infectious Disease Threats in the Twenty-First Century: Strengthening the Global Response'. *Frontiers in Immunology* 10. https://www.frontiersin.org/article/10.3389/fimmu.2019.00549.

Bohn, Mary Kathryn, Alexandra Hall, Lusia Sepiashvili, Benjamin Jung, Shannon Steele, and Khosrow Adeli. 2020. 'Pathophysiology of COVID-19: Mechanisms Underlying Disease Severity and Progression'. *Physiology* 35 (5): 288–301. https://doi.org/10.1152/physiol.00019.2020.

Bonferroni, C. 1936. 'Teoria Statistica Delle Classi e Calcolo Delle Probabilita'. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commericiali Di Firenze* 8: 3–62. https://ci.nii.ac.jp/naid/20001561442/.

Boos, Denni D., and L. A. Stefanski. 2013. 'M-Estimation (Estimating Equations)'. In *Essential Statistical Inference: Theory and Methods*, edited by Dennis D Boos and L. A Stefanski, 297–337. Springer Texts in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-4818-1_7.

Boysen, L., H. Rosenquist, J. T. Larsson, E. M. Nielsen, G. Sørensen, S. Nordentoft, and T. Hald. 2014. 'Source Attribution of Human Campylobacteriosis in Denmark'. *Epidemiology & Infection* 142 (8): 1599–1608. https://doi.org/10.1017/S0950268813002719.

Brake, Samuel James, Kathryn Barnsley, Wenying Lu, Kielan Darcy McAlinden, Mathew Suji Eapen, and Sukhwinder Singh Sohal. 2020. 'Smoking Upregulates Angiotensin-Converting Enzyme-2 Receptor: A Potential Adhesion Site for Novel Coronavirus SARS-CoV-2 (Covid-19)'. *Journal of Clinical Medicine* 9 (3): E841. https://doi.org/10.3390/jcm9030841.

Bramer, Max. 2013. 'Avoiding Overfitting of Decision Trees'. In *Principles of Data Mining*, edited by Max Bramer, 121–36. Undergraduate Topics in Computer Science. London: Springer. https://doi.org/10.1007/978-1-4471-4884-5_9.

Breiman, Leo. 2001a. 'Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)'. *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

———. 2001b. 'Random Forests'. *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brenner, Hermann, and Maria Blettner. 1997. 'Controlling for Continuous Confounders in Epidemiologic Research'. *Epidemiology* 8 (4): 429–34. https://www.jstor.org/stable/3702586.

Brotman, Daniel J., Esteban Walker, Michael S. Lauer, and Ralph G. O'Brien. 2005. 'In Search of Fewer Independent Risk Factors'. *Archives of Internal Medicine* 165 (2): 138–45. https://doi.org/10.1001/archinte.165.2.138.

Buchanan, Cody J., Andrew L. Webb, Steven K. Mutschall, Peter Kruczkiewicz, Dillon O. R. Barker, Benjamin M. Hetman, Victor P. J. Gannon, et al. 2017. 'A Genome-Wide Association Study to Identify Diagnostic Markers for Human Pathogenic Campylobacter Jejuni Strains'. *Frontiers in Microbiology* 8. https://www.frontiersin.org/article/10.3389/fmicb.2017.01224.

Bunnell, Katherine M., Tanayott Thaweethai, Colleen Buckless, Daniel J. Shinnick, Martin Torriani, Andrea S. Foulkes, and Miriam A. Bredella. 2021. 'Body Composition Predictors of Outcome in Patients with COVID-19'. *International Journal of Obesity* 45 (10): 2238–43. https://doi.org/10.1038/s41366-021-00907-1.

Burki, N. K., and R. W. Baker. 1984. 'Ventilatory Regulation in Eucapnic Morbid Obesity'. *The American Review of Respiratory Disease* 129 (4): 538–43.

Bushnell, Brian. 2014. 'BBMap: A Fast, Accurate, Splice-Aware Aligner'. LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner.

Butcher, James, Sabina Sarvan, Joseph S. Brunzelle, Jean-François Couture, and Alain Stintzi. 2012. 'Structure and Regulon of Campylobacter Jejuni Ferric Uptake Regulator Fur Define Apo-Fur Regulation'. *Proceedings of the National Academy of Sciences* 109 (25): 10047–52. https://doi.org/10.1073/pnas.1118321109.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. 'Statistics versus Machine Learning'. *Nature Methods* 15 (4): 233–34. https://doi.org/10.1038/nmeth.4642.

Bzdok, Danilo, Denis Engemann, and Bertrand Thirion. 2020. 'Inference and Prediction Diverge in Biomedicine'. *Patterns* 1 (8): 100119. https://doi.org/10.1016/j.patter.2020.100119.

Bzdok, Danilo, Martin Krzywinski, and Naomi Altman. 2017. 'Machine Learning: A Primer'. *Nature Methods* 14 (12): 1119–20. https://doi.org/10.1038/nmeth.4526.

Cabinet Office. 2020. 'New Rules on Staying at Home and Away from Others'. available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/883116/Staying_at_home_and_away_from_others__social_distancing_.pdf. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/883116/Staying_at_home_and_away_from_others__social_distancing_.pdf.

Cam, L. Le. 1990. 'Maximum Likelihood: An Introduction'. *International Statistical Review / Revue Internationale de Statistique* 58 (2): 153. https://doi.org/10.2307/1403464.

Cameron, Andrew, Emilisa Frirdich, Steven Huynh, Craig T. Parker, and Erin C. Gaynor. 2012. 'Hyperosmotic Stress Response of Campylobacter Jejuni'. *Journal of Bacteriology* 194 (22): 6116–30. https://doi.org/10.1128/JB.01409-12.

Cameron, Andrew, Steven Huynh, Nichollas E. Scott, Emilisa Frirdich, Dmitry Apel, Leonard J. Foster, Craig T. Parker, and Erin C. Gaynor. 2015. 'High-Frequency Variation of Purine Biosynthesis Genes Is a Mechanism of Success in Campylobacter Jejuni'. *MBio* 6 (5): e00612-15. https://doi.org/10.1128/mBio.00612-15.

Campbell, Michael J., and T. D. V. Swinscow. 2009. *Statistics at Square One*. Wiley.

Candon, Heather L., Brenda J. Allan, Cresson D. Fraley, and Erin C. Gaynor. 2007. 'Polyphosphate Kinase 1 Is a Pathogenesis Determinant in Campylobacter Jejuni'. *Journal of Bacteriology*, November. https://doi.org/10.1128/JB.01037-07.

Casella, George, and Roger L. Berger. 2021. *Statistical Inference*. Cengage Learning.

Chandrashekhar, Kshipra, Issmat I Kassem, Corey Nislow, Dharanesh Gangaiah, Rosario A Candelero-Rueda, and Gireesh Rajashekara. 2015. 'Transcriptome Analysis of

Campylobacter Jejuni Polyphosphate Kinase (*Ppk1* and *Ppk2*) Mutants'. *Virulence* 6 (8): 814–18. https://doi.org/10.1080/21505594.2015.1104449.

Chandrashekhar, Kshipra, Vishal Srivastava, Sunyoung Hwang, Byeonghwa Jeon, Sangryeol Ryu, and Gireesh Rajashekara. 2018. 'Transducer-Like Protein in Campylobacter Jejuni With a Role in Mediating Chemotaxis to Iron and Phosphate'. *Frontiers in Microbiology* 9. https://www.frontiersin.org/article/10.3389/fmicb.2018.02674.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie. 1987. 'A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation'. *Journal of Chronic Diseases* 40 (5): 373–83. https://doi.org/10.1016/0021-9681(87)90171-8.

Chen, Nanshan, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, et al. 2020. 'Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: A Descriptive Study'. *The Lancet* 395 (10223): 507–13. https://doi.org/10.1016/S0140-6736(20)30211-7.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. 'Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor'. *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chen, Xi, and Hemant Ishwaran. 2012. 'Random Forests for Genomic Data Analysis'. *Genomics* 99 (6): 323–29. https://doi.org/10.1016/j.ygeno.2012.04.003.

Clayton, David G. 2009. 'Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes'. *PLOS Genetics* 5 (7): e1000540. https://doi.org/10.1371/journal.pgen.1000540.

Clift, Ashley K., Adam von Ende, Pui San Tan, Hannah M. Sallis, Nicola Lindson, Carol A. C. Coupland, Marcus R. Munafò, Paul Aveyard, Julia Hippisley-Cox, and Jemma C. Hopewell. 2022. 'Smoking and COVID-19 Outcomes: An Observational and Mendelian Randomisation Study Using the UK Biobank Cohort'. *Thorax* 77 (1): 65–73. https://doi.org/10.1136/thoraxjnl-2021-217080.

Cody, Alison J., James E. Bray, Keith A. Jolley, Noel D. McCarthy, and Martin C. J. Maiden. 2017a. 'Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter Jejuni and C. Coli Human Disease Isolates'. *Journal of Clinical Microbiology* 55 (7): 2086–97. https://doi.org/10.1128/JCM.00080-17.

———. 2017b. 'Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter Jejuni and C. Coli Human Disease Isolates'.

Journal of Clinical Microbiology 55 (7): 2086–97. https://doi.org/10.1128/JCM.00080-17.

Cody, Alison J, Martin CJ Maiden, Norval JC Strachan, and Noel D McCarthy. 2019. 'A Systematic Review of Source Attribution of Human Campylobacteriosis Using Multilocus Sequence Typing'. *Eurosurveillance* 24 (43). https://doi.org/10.2807/1560-7917.ES.2019.24.43.1800696.

Cohen, Mitchell L. 2000. 'Changing Patterns of Infectious Disease'. *Nature* 406 (6797): 762–67. https://doi.org/10.1038/35021206.

Cortes, Corinna, and Vladimir Vapnik. 1995. 'Support-Vector Networks'. *Machine Learning* 20 (3): 273–97. https://doi.org/10.1007/BF00994018.

COVID-19 Host Genetics Initiative. 2021. 'Mapping the Human Genetic Architecture of COVID-19'. *Nature* 600 (7889): 472–77. https://doi.org/10.1038/s41586-021-03767-x.

Cox, David Roxbee, and D. V. Hinkley. 1974. *Theoretical Statistics*. Chapman and Hall.

Dabbah, Mohammad A., Angus B. Reed, Adam T. C. Booth, Arrash Yassaee, Aleksa Despotovic, Benjamin Klasmer, Emily Binning, et al. 2021. 'Machine Learning Approach to Dynamic Risk Modeling of Mortality in COVID-19: A UK Biobank Study'. *Scientific Reports* 11 (1): 16936. https://doi.org/10.1038/s41598-021-95136-x.

Dearlove, Bethany L., Alison J. Cody, Ben Pascoe, Guillaume Méric, Daniel J. Wilson, and Samuel K. Sheppard. 2016a. 'Rapid Host Switching in Generalist Campylobacter Strains Erodes the Signal for Tracing Human Infections'. *The ISME Journal* 10 (3): 721–29. https://doi.org/10.1038/ismej.2015.149.

———. 2016b. 'Rapid Host Switching in Generalist Campylobacter Strains Erodes the Signal for Tracing Human Infections'. *The ISME Journal* 10 (3): 721–29. https://doi.org/10.1038/ismej.2015.149.

Demortier, Luc. 2013. 'Interval Estimation', 42.

Deneke, Carlus, Robert Rentzsch, and Bernhard Y. Renard. 2017a. 'PaPrBaG: A Machine Learning Approach for the Detection of Novel Pathogens from NGS Data'. *Scientific Reports* 7 (January): 39194. https://doi.org/10.1038/srep39194.

———. 2017b. 'PaPrBaG: A Machine Learning Approach for the Detection of Novel Pathogens from NGS Data'. *Scientific Reports* 7 (January): 39194. https://doi.org/10.1038/srep39194.

Di Giannatale, Elisabetta, Giuliano Garofolo, Alessandra Alessiani, Guido Di Donato, Luca Candeloro, Walter Vencia, Lucia Decastelli, and Francesca Marotta. 2016. 'Tracing Back Clinical Campylobacter Jejuni in the Northwest of Italy and Assessing Their Potential Source'. *Frontiers in Microbiology* 7 (June). https://doi.org/10.3389/fmicb.2016.00887.

DiCiccio, Thomas J., and Bradley Efron. 1996. 'Bootstrap Confidence Intervals'. *Statistical Science* 11 (3): 189–228. https://doi.org/10.1214/ss/1032280214.

Didikoglu, Altug, Asri Maharani, Neil Pendleton, Maria Mercè Canal, and Antony Payton. 2021. 'Early Life Factors and COVID-19 Infection in England: A Prospective Analysis of UK Biobank Participants'. *Early Human Development* 155 (April): 105326. https://doi.org/10.1016/j.earlhumdev.2021.105326.

Dingle, K. E., F. M. Colles, D. R. Wareing, R. Ure, A. J. Fox, F. E. Bolton, H. J. Bootsma, R. J. Willems, R. Urwin, and M. C. Maiden. 2001. 'Multilocus Sequence Typing System for Campylobacter Jejuni'. *J. Clin. Microbiol.* 39 (1): 14–23. https://doi.org/10.1128/JCM.39.1.14-23.2001.

Docherty, Annemarie B., Ewen M. Harrison, Christopher A. Green, Hayley E. Hardwick, Riinu Pius, Lisa Norman, Karl A. Holden, et al. 2020. 'Features of 20 133 UK Patients in Hospital with Covid-19 Using the ISARIC WHO Clinical Characterisation Protocol: Prospective Observational Cohort Study'. *BMJ* 369 (May): m1985. https://doi.org/10.1136/bmj.m1985.

Doherty, Jean-François, Xuhong Chai, Laurie E. Cope, Daniela de Angeli Dutra, Marin Milotic, Steven Ni, Eunji Park, and Antoine Filion. 2021. 'The Rise of Big Data in Disease Ecology'. *Trends in Parasitology* 37 (12): 1034–37. https://doi.org/10.1016/j.pt.2021.09.003.

Dórea, Fernanda C., and Crawford W. Revie. 2021. 'Data-Driven Surveillance: Effective Collection, Integration, and Interpretation of Data to Support Decision Making'. *Frontiers in Veterinary Science* 8. https://www.frontiersin.org/article/10.3389/fvets.2021.633977.

Douglas, L. 2012. 'The Importance of Big Data: A Definition. Gartner'. *Online: Http://Www. Gartner. Com/ResId* 2057415 (21.06): 2012.

Earle, Sarah G., Mariya Lobanovska, Hayley Lavender, Changyan Tang, Rachel M. Exley, Elisa Ramos-Sevillano, Douglas F. Browning, et al. 2021. 'Genome-Wide Association Studies Reveal the Role of Polymorphisms Affecting Factor H Binding Protein Expression in Host Invasion by Neisseria Meningitidis'. *PLOS Pathogens* 17 (10): e1009992. https://doi.org/10.1371/journal.ppat.1009992.

Earle, Sarah G., Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N. Claire Gordon, Timothy M. Walker, Chris C. A. Spencer, et al. 2016. 'Identifying Lineage Effects When Controlling for Population Structure Improves Power in Bacterial Association Studies'. *Nature Microbiology* 1 (5): 1–8. https://doi.org/10.1038/nmicrobiol.2016.41.

Eberly, Lynn E., and George Casella. 2003. 'Estimating Bayesian Credible Intervals'. *Journal of Statistical Planning and Inference*, Special issue II: Model Selection, Model Diagnostics, Empirical B ayes and Hierarchical Bayes, 112 (1): 115–32. https://doi.org/10.1016/S0378-3758(02)00327-0.

Eckhardt, Manon, Judd F. Hultquist, Robyn M. Kaake, Ruth Hüttenhain, and Nevan J. Krogan. 2020. 'A Systems Approach to Infectious Disease'. *Nature Reviews Genetics* 21 (6): 339–54. https://doi.org/10.1038/s41576-020-0212-5.

Efron, Bradley, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.

Ehrenstein, Vera, Henrik Nielsen, Alma B Pedersen, Søren P Johnsen, and Lars Pedersen. 2017. 'Clinical Epidemiology in the Era of Big Data: New Opportunities, Familiar Challenges'. *Clinical Epidemiology* 9 (April): 245–50. https://doi.org/10.2147/CLEP.S129779.

Elliott, Joshua, Barbara Bodinier, Matthew Whitaker, Cyrille Delpierre, Roel Vermeulen, Ioanna Tzoulaki, Paul Elliott, and Marc Chadeau-Hyam. 2021. 'COVID-19 Mortality in the UK Biobank Cohort: Revisiting and Evaluating Risk Factors'. *European Journal of Epidemiology* 36 (3): 299–309. https://doi.org/10.1007/s10654-021-00722-y.

Epping, Lennard, Birgit Walther, Rosario M. Piro, Marie-Theres Knüver, Charlotte Huber, Andrea Thürmer, Antje Flieger, et al. 2021. 'Genome-Wide Insights into Population Structure and Host Specificity of Campylobacter Jejuni'. *Scientific Reports* 11 (1): 10358. https://doi.org/10.1038/s41598-021-89683-6.

European Food Safety Authority and European Centre for Disease Prevention and Control. 2020. 'The European Union Summary Report on Antimicrobial Resistance in Zoonotic and Indicator Bacteria from Humans, Animals and Food in 2017/2018'. *EFSA Journal* 18 (3). https://doi.org/10.2903/j.efsa.2020.6007.

Fatima, Yaqoot, Romola S. Bucks, Abdullah A. Mamun, Isabelle Skinner, Ivana Rosenzweig, Guy Leschziner, and Timothy C. Skinner. 2021. 'Shift Work Is Associated with Increased Risk of COVID-19: Findings from the UK Biobank Cohort'. *Journal of Sleep Research* 30 (5): e13326. https://doi.org/10.1111/jsr.13326.

Favre, Guillaume, Kevin Legueult, Christian Pradier, Charles Raffaelli, Carole Ichai, Antonio Iannelli, Alban Redheuil, Olivier Lucidarme, and Vincent Esnault. 2021. 'Visceral Fat Is Associated to the Severity of COVID-19'. *Metabolism* 115 (February): 154440. https://doi.org/10.1016/j.metabol.2020.154440.

Feldgarden, Michael, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G. Frye, Julie Haendiges, Daniel H. Haft, Maria Hoffmann, et al. 2021. 'AMRFinderPlus and the Reference Gene Catalog Facilitate Examination of the Genomic Links among Antimicrobial Resistance, Stress Response, and Virulence'. *Scientific Reports* 11 (1): 12728. https://doi.org/10.1038/s41598-021-91456-0.

Ferrari, Elisa, Alessandra Retico, and Davide Bacciu. 2020. 'Measuring the Effects of Confounders in Medical Supervised Classification Problems: The Confounding Index (CI)'. *Artificial Intelligence in Medicine* 103 (March): 101804. https://doi.org/10.1016/j.artmed.2020.101804.

Fix, Evelyn, and Jr Hodges. 1951. 'Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties'. CALIFORNIA UNIV BERKELEY. https://apps.dtic.mil/sti/citations/ADA800276.

Földi, Mária, Nelli Farkas, Szabolcs Kiss, Fanni Dembrovszky, Zsolt Szakács, Márta Balaskó, Bálint Erőss, Péter Hegyi, and Andrea Szentesi. 2021. 'Visceral Adiposity Elevates the Risk of Critical Condition in COVID-19: A Systematic Review and Meta-Analysis'. *Obesity* 29 (3): 521–28. https://doi.org/10.1002/oby.23096.

Fontana, Luigi, J. Christopher Eagon, Maria E. Trujillo, Philipp E. Scherer, and Samuel Klein. 2007. 'Visceral Fat Adipokine Secretion Is Associated with Systemic Inflammation in Obese Humans'. *Diabetes* 56 (4): 1010–13. https://doi.org/10.2337/db06-1656.

Freuer, Dennis, Jakob Linseisen, and Christa Meisinger. 2021. 'Impact of Body Composition on COVID-19 Susceptibility and Severity: A Two-Sample Multivariable Mendelian Randomization Study'. *Metabolism* 118: 154732. https://doi.org/10.1016/j.metabol.2021.154732.

Friedman, Jerome H. 2001. 'Greedy Function Approximation: A Gradient Boosting Machine.' *The Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Furukawa, Nathan W., John T. Brooks, and Jeremy Sobel. 2020. 'Evidence Supporting Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 While Presymptomatic or Asymptomatic'. *Emerging Infectious Diseases* 26 (7). https://doi.org/10.3201/eid2607.201595.

Gao, Min, Qin Wang, Carmen Piernas, Nerys M. Astbury, Susan A. Jebb, Michael V. Holmes, and Paul Aveyard. 2022. 'Associations between Body Composition, Fat Distribution and Metabolic Consequences of Excess Adiposity with Severe COVID-19 Outcomes: Observational Study and Mendelian Randomisation Analysis'. *International Journal of Obesity*, January, 1–8. https://doi.org/10.1038/s41366-021-01054-3.

Gaynor, Erin C., Derek H. Wells, Joanna K. MacKichan, and Stanley Falkow. 2005. 'The Campylobacter Jejuni Stringent Response Controls Specific Stress Survival and Virulence-Associated Phenotypes'. *Molecular Microbiology* 56 (1): 8–27. https://doi.org/10.1111/j.1365-2958.2005.04525.x.

Gilbert, Maarten J., William G. Miller, Emma Yee, Aldert L. Zomer, Linda van der Graaf-van Bloois, Collette Fitzgerald, Ken J. Forbes, et al. 2016. 'Comparative Genomics of Campylobacter Fetus from Reptiles and Mammals Reveals Divergent Evolution in Host-Associated Lineages'. *Genome Biology and Evolution* 8 (6): 2006–19. https://doi.org/10.1093/gbe/evw146.

Gilks, W. R., S. Richardson, and David Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. CRC Press.

Gillies, Clare L, Alex V Rowlands, Cameron Razieh, Vahé Nafilyan, Yogini Chudasama, Nazrul Islam, Francesco Zaccardi, et al. 2022. 'Association between Household Size and COVID-19: A UK Biobank Observational Study'. *Journal of the Royal Society of Medicine*, February, 01410768211073923. https://doi.org/10.1177/01410768211073923.

Gilmour, M. W., M. Graham, A. Reimer, and G. Van Domselaar. 2013. 'Public Health Genomics and the New Molecular Epidemiology of Bacterial Pathogens'. *Public Health Genomics* 16 (1–2): 25–30. https://doi.org/10.1159/000342709.

Giraud, Antoine, Ivan Matic, Olivier Tenaillon, Antonio Clara, Miroslav Radman, Michel Fons, and François Taddei. 2001. 'Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut'. *Science*, March. https://doi.org/10.1126/science.1056421.

Greene, Anna C., Kristine A. Giffin, Casey S. Greene, and Jason H. Moore. 2016. 'Adapting Bioinformatics Curricula for Big Data'. *Briefings in Bioinformatics* 17 (1): 43–50. https://doi.org/10.1093/bib/bbv018.

Greenland, Sander, Judea Pearl, and James M. Robins. 1999. 'Confounding and Collapsibility in Causal Inference'. *Statistical Science* 14 (1): 29–46. https://doi.org/10.1214/ss/1009211805.

Griffith, Gareth J., Tim T. Morris, Matthew J. Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, et al. 2020. 'Collider Bias Undermines Our Understanding of COVID-19 Disease Risk and Severity'. *Nature Communications* 11 (1): 5749. https://doi.org/10.1038/s41467-020-19478-2.

Guhathakurata, Soham, Souvik Kundu, Arpita Chakraborty, and Jyoti Sekhar Banerjee. 2021. 'A Novel Approach to Predict COVID-19 Using Support Vector Machine'. *Data Science for COVID-19*, 351–64. https://doi.org/10.1016/B978-0-12-824536-1.00014-9.

Guyatt, G, R Jaeschke, N Heddle, D Cook, H Shannon, and S Walter. 1995. 'Basic Statistics for Clinicians: 1. Hypothesis Testing.' *CMAJ: Canadian Medical Association Journal* 152 (1): 27–32. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1337490/.

Haldar, Swapan Kumar. 2018. 'Chapter 9 - Statistical and Geostatistical Applications in Geology'. In *Mineral Exploration (Second Edition)*, edited by Swapan Kumar Haldar, 167–94. Elsevier. https://doi.org/10.1016/B978-0-12-814022-2.00009-5.

Hall, Alastair R. 2004. *Generalized Method of Moments*. OUP Oxford.

Halsey, Lewis G. 2019. 'The Reign of the P-Value Is over: What Alternative Analyses Could We Employ to Fill the Power Vacuum?' *Biology Letters* 15 (5): 20190174. https://doi.org/10.1098/rsbl.2019.0174.

Hamer, Mark, Mika Kivimäki, Catharine R. Gale, and G. David Batty. 2020. 'Lifestyle Risk Factors, Inflammatory Mechanisms, and COVID-19 Hospitalization: A Community-Based

Cohort Study of 387,109 Adults in UK'. *Brain, Behavior, and Immunity* 87 (July): 184–87. https://doi.org/10.1016/j.bbi.2020.05.059.

Han, Lefei, Jinjun Ran, Yim-Wah Mak, Lorna Kwai-Ping Suen, Paul H. Lee, Joseph Sriyal Malik Peiris, and Lin Yang. 2019. 'Smoking and Influenza-Associated Morbidity and Mortality: A Systematic Review and Meta-Analysis'. *Epidemiology (Cambridge, Mass.)* 30 (3): 405–17. https://doi.org/10.1097/EDE.0000000000000984.

Haseli, Sara, Nastaran Khalili, Mehrdad Bakhshayeshkaram, Morteza Sanei Taheri, and Yashar Moharramzad. 2020. 'Lobar Distribution of COVID-19 Pneumonia Based on Chest Computed Tomography Findings; A Retrospective Study'. *Archives of Academic Emergency Medicine* 8 (1): e55. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7212068/.

Hastie, Claire E., Jill P. Pell, and Naveed Sattar. 2021. 'Vitamin D and COVID-19 Infection and Mortality in UK Biobank'. *European Journal of Nutrition* 60 (1): 545–48. https://doi.org/10.1007/s00394-020-02372-4.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. 'The Elements of Statistical Learning – Data Mining, Inference, and Prediction'.

Hautakangas, Heidi, Bendik S. Winsvold, Sanni E. Ruotsalainen, Gyda Bjornsdottir, Aster V. E. Harder, Lisette J. A. Kogelman, Laurent F. Thomas, et al. 2022. 'Genome-Wide Analysis of 102,084 Migraine Cases Identifies 123 Risk Loci and Subtype-Specific Risk Alleles'. *Nature Genetics* 54 (2): 152–60. https://doi.org/10.1038/s41588-021-00990-0.

Hazarie, Surendra, David Soriano-Paños, Alex Arenas, Jesús Gómez-Gardeñes, and Gourab Ghoshal. 2021. 'Interplay between Population Density and Mobility in Determining the Spread of Epidemics in Cities'. *Communications Physics* 4 (1): 1–10. https://doi.org/10.1038/s42005-021-00679-0.

Hazra, Avijit. 2017. 'Using the Confidence Interval Confidently'. *Journal of Thoracic Disease* 9 (10): 4125–30. https://doi.org/10.21037/jtd.2017.09.14.

He, Sunyue, Jie Yang, Xiaoyong Li, Hongxia Gu, Qing Su, and Li Qin. 2021. 'Visceral Adiposity Index Is Associated with Lung Function Impairment: A Population-Based Study'. *Respiratory Research* 22 (1): 2. https://doi.org/10.1186/s12931-020-01599-3.

Hedge, Jessica, and Daniel J. Wilson. 2014. 'Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not'. *MBio* 5 (6). https://doi.org/10.1128/mBio.02158-14.

Hermans, David, Kim Van Deun, An Martel, Filip Van Immerseel, Winy Messens, Marc Heyndrickx, Freddy Haesebrouck, and Frank Pasmans. 2011. 'Colonization Factors of Campylobacter Jejuni in the Chicken Gut'. *Veterinary Research* 42 (1): 82. https://doi.org/10.1186/1297-9716-42-82.

Hilton, Bridget, Daniel Wilson, Anne-Marie O'Connell, Dean Ironmonger, Justine K. Rudkin, Naomi Allen, Isabel Oliver, and David Wyllie. 2020. 'Microbial Isolation in English Participants in the UK Biobank Cohort: Comparison with the General Population'. medRxiv. https://doi.org/10.1101/2020.03.18.20038281.

Hinne, Max, Quentin F. Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. 2020. 'A Conceptual Introduction to Bayesian Model Averaging'. *Advances in Methods and Practices in Psychological Science* 3 (2): 200–215. https://doi.org/10.1177/2515245919898657.

Ho, Frederick K., Carlos A. Celis-Morales, Stuart R. Gray, S. Vittal Katikireddi, Claire L. Niedzwiedz, Claire Hastie, Lyn D. Ferguson, et al. 2020. 'Modifiable and Non-Modifiable Risk Factors for COVID-19, and Comparison to Risk Factors for Influenza and Pneumonia: Results from a UK Biobank Prospective Cohort Study'. *BMJ Open* 10 (11): e040402. https://doi.org/10.1136/bmjopen-2020-040402.

Ho, Frederick K., Fanny Petermann-Rocha, Stuart R. Gray, Bhautesh D. Jani, S. Vittal Katikireddi, Claire L. Niedzwiedz, Hamish Foster, et al. 2020. 'Is Older Age Associated with COVID-19 Mortality in the Absence of Other Risk Factors? General Population Cohort Study of 470,034 Participants'. *PLOS ONE* 15 (11): e0241824. https://doi.org/10.1371/journal.pone.0241824.

Hoerl, Arthur E., and Robert W. Kennard. 1970. 'Ridge Regression: Biased Estimation for Nonorthogonal Problems'. *Technometrics* 12 (1): 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Hoffman, Sharona, and Andy Podgurski. 2013. 'Big Bad Data: Law, Public Health, and Biomedical Databases'. *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics* 41 Suppl 1 (March): 56–60. https://doi.org/10.1111/jlme.12040.

Hofreuter, Dirk, Veronica Novik, and Jorge E. Galán. 2008. 'Metabolic Diversity in Campylobacter Jejuni Enhances Specific Tissue Colonization'. *Cell Host & Microbe* 4 (5): 425–33. https://doi.org/10.1016/j.chom.2008.10.002.

Hoijtink, Herbert, Joris Mulder, Caspar van Lissa, and Xin Gu. 2019. 'A Tutorial on Testing Hypotheses Using the Bayes Factor'. *Psychological Methods* 24 (5): 539–56. https://doi.org/10.1037/met0000201.

Holz, Hilary J., and Murray H. Loew. 1994. 'Relative Feature Importance: A Classifier-Independent Approach to Feature Selection'. In *Machine Intelligence and Pattern Recognition*, edited by Edzard S. Gelsema and Laveen S. Kanal, 16:473–87. Pattern Recognition in Practice IV. North-Holland. https://doi.org/10.1016/B978-0-444-81892-8.50046-8.

Hopfield, J J. 1982. ‘Neural Networks and Physical Systems with Emergent Collective Computational Abilities.’ *Proceedings of the National Academy of Sciences* 79 (8): 2554–58. https://doi.org/10.1073/pnas.79.8.2554.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. ‘Unbiased Recursive Partitioning: A Conditional Inference Framework’. *Journal of Computational and Graphical Statistics* 15 (3): 651–74. https://doi.org/10.1198/106186006X133933.

Hu, Donglei, and Elad Ziv. 2008. ‘Confounding in Genetic Association Studies and Its Solutions’. *Methods in Molecular Biology (Clifton, N.J.)* 448: 31–39. https://doi.org/10.1007/978-1-59745-205-2_3.

Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, et al. 2020. ‘Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China’. *The Lancet* 395 (10223): 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

Huettner, Frank, and Marco Sunder. 2012. ‘Axiomatic Arguments for Decomposing Goodness of Fit According to Shapley and Owen Values’. *Electronic Journal of Statistics* 6 (none): 1239–50. https://doi.org/10.1214/12-EJS710.

Humer, E., C. Schwarz, and K. Schedle. 2015. ‘Phytate in Pig and Poultry Nutrition’. *Journal of Animal Physiology and Animal Nutrition* 99 (4): 605–25. https://doi.org/10.1111/jpn.12258.

Hunter, John D. 2007. ‘Matplotlib: A 2D Graphics Environment’. *Computing in Science Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Huttunen, R., T. Heikkinen, and J. Syrjänen. 2011. ‘Smoking and the Outcome of Infection’. *Journal of Internal Medicine* 269 (3): 258–69. https://doi.org/10.1111/j.1365-2796.2010.02332.x.

Institute of Environmental Science and Research Ltd. 2007. ‘Notifiable and Other Diseases in New Zealand: Annual Report 2006’. *Porirua (NZ): The Institute*.

Iwashyna, Theodore J., and Vincent Liu. 2014. ‘What’s So Different about Big Data?. A Primer for Clinicians Trained to Think Epidemiologically’. *Annals of the American Thoracic Society* 11 (7): 1130–35. https://doi.org/10.1513/AnnalsATS.201405-185AS.

Jagannathan, Aparna, Chrystala Constantinidou, and Charles W. Penn. 2001. ‘Roles of RpoN, FliA,and FlgR in Expression of Flagella InCampylobacter Jejuni’. *Journal of Bacteriology*, May. https://doi.org/10.1128/JB.183.9.2937-2942.2001.

Jakobsdottir, Johanna, Michael B. Gorin, Yvette P. Conley, Robert E. Ferrell, and Daniel E. Weeks. 2009. ‘Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers’. *PLOS Genetics* 5 (2): e1000337. https://doi.org/10.1371/journal.pgen.1000337.

Janssens, A. Cecile J. W., and Cornelia M. van Duijn. 2008. 'Genome-Based Prediction of Common Diseases: Advances and Prospects'. *Human Molecular Genetics* 17 (R2): R166-173. https://doi.org/10.1093/hmg/ddn250.

Jian, Zhongyu, Menghua Wang, Xi Jin, and Xin Wei. 2021. 'Genetically Predicted Higher Educational Attainment Decreases the Risk of COVID-19 Susceptibility and Severity: A Mendelian Randomization Study'. *Frontiers in Public Health* 9 (December): 731962. https://doi.org/10.3389/fpubh.2021.731962.

Jing, Qin-Long, Ming-Jin Liu, Zhou-Bin Zhang, Li-Qun Fang, Jun Yuan, An-Ran Zhang, Natalie E. Dean, et al. 2020. 'Household Secondary Attack Rate of COVID-19 and Associated Determinants in Guangzhou, China: A Retrospective Cohort Study'. *The Lancet Infectious Diseases* 20 (10): 1141–50. https://doi.org/10.1016/S1473-3099(20)30471-0.

John, George H., and Pat Langley. 2013. 'Estimating Continuous Distributions in Bayesian Classifiers'. *ArXiv:1302.4964 [Cs, Stat]*, February. http://arxiv.org/abs/1302.4964.

Jolley, Keith A., James E. Bray, and Martin C. J. Maiden. 2018. 'Open-Access Bacterial Population Genomics: BIGSdb Software, the PubMLST.Org Website and Their Applications'. *Wellcome Open Research* 3 (September): 124. https://doi.org/10.12688/wellcomeopenres.14826.1.

Jolley, Keith A., and Martin CJ Maiden. 2010. 'BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level'. *BMC Bioinformatics* 11 (1): 595. https://doi.org/10.1186/1471-2105-11-595.

Kaakoush, Nadeem O., Natalia Castaño-Rodríguez, Hazel M. Mitchell, and Si Ming Man. 2015. 'Global Epidemiology of Campylobacter Infection'. *Clinical Microbiology Reviews* 28 (3): 687–720. https://doi.org/10.1128/CMR.00006-15.

Kaler, Avjinder S., and Larry C. Purcell. 2019. 'Estimation of a Significance Threshold for Genome-Wide Association Studies'. *BMC Genomics* 20 (1): 618. https://doi.org/10.1186/s12864-019-5992-7.

Kashyap, Hirak, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. 2016. 'Big Data Analytics in Bioinformatics: Architectures, Techniques, Tools and Issues'. *Network Modeling Analysis in Health Informatics and Bioinformatics* 5 (1): 28. https://doi.org/10.1007/s13721-016-0135-4.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform'. *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'. In

Proceedings of the 31st International Conference on Neural Information Processing Systems, 3149–57. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.

Khalangot, Mykola, Nadiia Sheichenko, Vitaly Gurianov, Viola Vlasenko, Yulia Kurinna, Oksana Samson, and Mykola Tronko. 2022. 'Relationship between Hyperglycemia, Waist Circumference, and the Course of COVID-19: Mortality Risk Assessment'. *Experimental Biology and Medicine (Maywood, N.J.)* 247 (3): 200–206. https://doi.org/10.1177/15353702211054452.

Khaodhiar, L., K. C. McCowen, and G. L. Blackburn. 1999. 'Obesity and Its Comorbid Conditions'. *Clinical Cornerstone* 2 (3): 17–31. https://doi.org/10.1016/s1098-3597(99)90002-9.

Khoury, Muin J., and John P. A. Ioannidis. 2014. 'Big Data Meets Public Health'. *Science (New York, N.Y.)* 346 (6213): 1054–55. https://doi.org/10.1126/science.aaa2709.

Kingma, Diederik P., and Jimmy Ba. 2014. 'Adam: A Method for Stochastic Optimization'. *ArXiv:1412.6980 [Cs]*, December. http://arxiv.org/abs/1412.6980.

Kirk, Karina Frahm, Guillaume Méric, Hans Linde Nielsen, Ben Pascoe, Samuel K. Sheppard, Ole Thorlacius-Ussing, and Henrik Nielsen. 2018. 'Molecular Epidemiology and Comparative Genomics of Campylobacter Concisus Strains from Saliva, Faeces and Gut Mucosal Biopsies in Inflammatory Bowel Disease'. *Scientific Reports* 8 (1): 1902. https://doi.org/10.1038/s41598-018-20135-4.

Kittl, Sonja, Gerald Heckel, Bożena M. Korczak, and Peter Kuhnert. 2013. 'Source Attribution of Human Campylobacter Isolates by MLST and Fla-Typing and Association of Genotypes with Quinolone Resistance'. *PLOS ONE* 8 (11): e81796. https://doi.org/10.1371/journal.pone.0081796.

Koch, Karl-Rudolf. 1999. *Parameter Estimation and Hypothesis Testing in Linear Models*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-03976-2.

Kolin, David A., Scott Kulm, Paul J. Christos, and Olivier Elemento. 2020. 'Clinical, Regional, and Genetic Characteristics of Covid-19 Patients from UK Biobank'. *PloS One* 15 (11): e0241264. https://doi.org/10.1371/journal.pone.0241264.

Kolmogorov, Mikhail, Brian Raney, Benedict Paten, and Son Pham. 2014. 'Ragout—a Reference-Assisted Assembly Tool for Bacterial Genomes'. *Bioinformatics* 30 (12): i302. https://doi.org/10.1093/bioinformatics/btu280.

Kotsiantis, S. B., I. D. Zaharakis, and P. E. Pintelas. 2006. 'Machine Learning: A Review of Classification and Combining Techniques'. *Artif Intell Rev* 26 (3): 159–90. https://doi.org/10.1007/s10462-007-9052-3.

Kraemer, Helena Chmura, Eric Stice, Alan Kazdin, David Offord, and David Kupfer. 2001. 'How Do Risk Factors Work Together? Mediators, Moderators, and Independent,

Overlapping, and Proxy Risk Factors'. *American Journal of Psychiatry* 158 (6): 848–56. https://doi.org/10.1176/appi.ajp.158.6.848.

Kwan, Patrick S. L., Andrew Birtles, Frederick J. Bolton, Nigel P. French, Susan E. Robinson, Lynne S. Newbold, Mathew Upton, and Andrew J. Fox. 2008. 'Longitudinal Study of the Molecular Epidemiology of Campylobacter Jejuni in Cattle on Dairy Farms'. *Applied and Environmental Microbiology* 74 (12): 3626–33. https://doi.org/10.1128/AEM.01669-07.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. 'Initial Sequencing and Analysis of the Human Genome'. *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Langmead, Ben, and Steven L Salzberg. 2012. 'Fast Gapped-Read Alignment with Bowtie 2'. *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Larvin, H., S. Wilmott, J. Kang, V. R. Aggarwal, S. Pavitt, and J. Wu. 2021. 'Additive Effect of Periodontal Disease and Obesity on COVID-19 Outcomes'. *Journal of Dental Research* 100 (11): 1228–35. https://doi.org/10.1177/00220345211029638.

Larvin, Harriet, Sheryl Wilmott, Jianhua Wu, and Jing Kang. 2020. 'The Impact of Periodontal Disease on Hospital Admission and Mortality During COVID-19 Pandemic'. *Frontiers in Medicine* 7. https://www.frontiersin.org/article/10.3389/fmed.2020.604980.

Lashley, Felissa R. 2006. 'Emerging Infectious Diseases at the Beginning of the 21st Century'. *Online Journal of Issues in Nursing* 11 (1): 2.

Lassale, Camille, Bamba Gaye, Mark Hamer, Catharine R. Gale, and G David Batty. 2020. 'Ethnic Disparities in Hospitalisation for COVID-19 in England: The Role of Socioeconomic Factors, Mental Health, and Inflammatory and pro-Inflammatory Factors in a Community-Based Cohort Study'. *Brain, Behavior, and Immunity* 88 (August): 44–49. https://doi.org/10.1016/j.bbi.2020.05.074.

Lassale, Camille, Mark Hamer, Álvaro Hernáez, Catharine R. Gale, and G. David Batty. 2021. 'Association of Pre-Pandemic High-Density Lipoprotein Cholesterol with Risk of COVID-19 Hospitalisation and Death: The UK Biobank Cohort Study'. *MedRxiv*, February, 2021.01.20.21250152. https://doi.org/10.1101/2021.01.20.21250152.

Laxminarayan, Ramanan. 2022. 'The Overlooked Pandemic of Antimicrobial Resistance'. *The Lancet* 399 (10325): 606–7. https://doi.org/10.1016/S0140-6736(22)00087-3.

Lee, Shing Fung, Maja Nikšić, Bernard Rachet, Maria-Jose Sanchez, and Miguel Angel Luque-Fernandez. 2021. 'Socioeconomic Inequalities and Ethnicity Are Associated with a Positive COVID-19 Test among Cancer Patients in the UK Biobank Cohort'. *Cancers* 13 (7): 1514. https://doi.org/10.3390/cancers13071514.

Lees, John A., T. Tien Mai, Marco Galardini, Nicole E. Wheeler, Samuel T. Horsfield, Julian Parkhill, and Jukka Corander. 2020. 'Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions'. *MBio* 11 (4). https://doi.org/10.1128/mBio.01344-20.

Lehrer, Steven, and Peter Rheinstein. 2021a. 'Eyeglasses Reduce Risk of COVID-19 Infection'. *In Vivo (Athens, Greece)* 35 (3): 1581–82. https://doi.org/10.21873/invivo.12414.

Lehrer, Steven, and Peter H. Rheinstein. 2021b. 'Common Drugs, Vitamins, Nutritional Supplements and COVID-19 Mortality'. *International Journal of Functional Nutrition* 2 (1): 1–5. https://doi.org/10.3892/ijfn.2021.14.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. 2011. 'The Sequence Read Archive'. *Nucleic Acids Research* 39 (Database issue): D19–21. https://doi.org/10.1093/nar/gkq1019.

Leon-Kempis, Maria del Rocio, Edward Guccione, Francis Mulholland, Michael P. Williamson, and David J. Kelly. 2006. 'The Campylobacter Jejuni PEB1a Adhesin Is an Aspartate/Glutamate-Binding Protein of an ABC Transporter Essential for Microaerobic Growth on Dicarboxylic Amino Acids'. *Molecular Microbiology* 60 (5): 1262–75. https://doi.org/10.1111/j.1365-2958.2006.05168.x.

Levy, Roger. 2012. 'Probabilistic Models in the Study of Language', 274.

Li, Fudong, Yi Shen, Duo Lv, Junfen Lin, Biyao Liu, Fan He, and Zhen Wang. 2020. 'A Bayesian Classification Model for Discriminating Common Infectious Diseases in Zhejiang Province, China'. *Medicine* 99 (8): e19218. https://doi.org/10.1097/MD.0000000000019218.

Li, Gloria Hoi-Yee, Stanley Kam-Ki Lam, Ian Chi-Kei Wong, Jody Kwok-Pui Chu, and Ching-Lung Cheung. 2021. 'Education Attainment, Intelligence and COVID-19: A Mendelian Randomization Study'. *Journal of Clinical Medicine* 10 (21): 4870. https://doi.org/10.3390/jcm10214870.

Li, Guoqiang, and Peifeng Niu. 2013. 'An Enhanced Extreme Learning Machine Based on Ridge Regression for Regression'. *Neural Computing and Applications* 22 (3): 803–10. https://doi.org/10.1007/s00521-011-0771-7.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Jiuling, Aowen Tian, Haoxue Zhu, Lanlan Chen, Jianping Wen, Wanqing Liu, and Peng Chen. 2022. 'Mendelian Randomization Analysis Reveals No Causal Relationship Between Nonalcoholic Fatty Liver Disease and Severe COVID-19'. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American*

Gastroenterological Association, February, S1542-3565(22)00104-5. https://doi.org/10.1016/j.cgh.2022.01.045.

Li, Mujin, Honghui Xu, and Yong Deng. 2019. 'Evidential Decision Tree Based on Belief Entropy'. *Entropy* 21 (9): 897. https://doi.org/10.3390/e21090897.

Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, et al. 2020. 'Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia'. *New England Journal of Medicine* 382 (13): 1199–1207. https://doi.org/10.1056/NEJMoa2001316.

Li, Shu, Zhi Cao, Hongxi Yang, Yuan Zhang, Fusheng Xu, and Yaogang Wang. 2021. 'Metabolic Healthy Obesity, Vitamin D Status, and Risk of COVID-19'. *Aging and Disease* 12 (1): 61–71. https://doi.org/10.14336/AD.2020.1108.

Li, Weizhong, and Adam Godzik. 2006. 'Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences'. *Bioinformatics* 22 (13): 1658–59. https://doi.org/10.1093/bioinformatics/btl158.

Li, Yu, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. 2019. 'Deep Learning in Bioinformatics: Introduction, Application, and Perspective in the Big Data Era'. *Methods*, Deep Learning in Bioinformatics, 166 (August): 4–21. https://doi.org/10.1016/j.ymeth.2019.04.008.

Libbrecht, Maxwell W., and William Stafford Noble. 2015. 'Machine Learning Applications in Genetics and Genomics'. *Nature Reviews Genetics* 16 (6): 321–32. https://doi.org/10.1038/nrg3920.

Lin, Ann E., Kirsten Krastel, Rhonda I. Hobb, Stuart A. Thompson, Dennis G. Cvitkovitch, and Erin C. Gaynor. 2009. 'Atypical Roles for Campylobacter Jejuni Amino Acid ATP Binding Cassette Transporter Components PaqP and PaqQ in Bacterial Stress Tolerance and Pathogen-Host Cell Dynamics'. *Infection and Immunity*, November. https://doi.org/10.1128/IAI.00571-08.

Liu, Jia, Daniel J. Nordman, and William Q. Meeker. 2016. 'The Number of MCMC Draws Needed to Compute Bayesian Credible Bounds'. *The American Statistician* 70 (3): 275–84. https://doi.org/10.1080/00031305.2016.1158738.

Lo, Adeline, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. 2015. 'Why Significant Variables Aren't Automatically Good Predictors'. *Proceedings of the National Academy of Sciences* 112 (45): 13892–97. https://doi.org/10.1073/pnas.1518285112.

Lockhart, Sam M., and Stephen O'Rahilly. 2020. 'When Two Pandemics Meet: Why Is Obesity Associated with Increased COVID-19 Mortality?' *Med (New York, N.y.)* 1 (1): 33–42. https://doi.org/10.1016/j.medj.2020.06.005.

Lodge, Caroline J., Alice Doherty, Dinh S. Bui, Raisa Cassim, Adrian J. Lowe, Alvar Agusti, Melissa A. Russell, and Shyamali C. Dharmage. 2021. 'Is Asthma Associated with COVID-

19 Infection? A UK Biobank Analysis'. *ERJ Open Research* 7 (4). https://doi.org/10.1183/23120541.00309-2021.

Loh, Wei-yin, and Yu-shan Shih. 1997. 'Split Selection Methods for Classification Trees'. *Statistica Sinica*, 815–40.

Luo, Juhua, Allison Hodge, Michael Hendryx, and Julie E. Byles. 2020. 'Age of Obesity Onset, Cumulative Obesity Exposure over Early Adulthood and Risk of Type 2 Diabetes'. *Diabetologia* 63 (3): 519–27. https://doi.org/10.1007/s00125-019-05058-7.

Luo, Naidan, Sonia Pereira, Orhan Sahin, Jun Lin, Shouxiong Huang, Linda Michel, and Qijing Zhang. 2005. 'Enhanced in Vivo Fitness of Fluoroquinolone-Resistant Campylobacter Jejuni in the Absence of Antibiotic Selection Pressure'. *Proceedings of the National Academy of Sciences* 102 (3): 541–46. https://doi.org/10.1073/pnas.0408966102.

Lupolova, Nadejda, Tim J. Dallman, Nicola J. Holden, and David L. Gally. 2017. 'Patchy Promiscuity: Machine Learning Applied to Predict the Host Specificity of Salmonella Enterica and Escherichia Coli'. *Microbial Genomics* 3 (10). https://doi.org/10.1099/mgen.0.000135.

Ma, Hao, Tao Zhou, Yoriko Heianza, and Lu Qi. 2021. 'Habitual Use of Vitamin D Supplements and Risk of Coronavirus Disease 2019 (COVID-19) Infection: A Prospective Study in UK Biobank'. *The American Journal of Clinical Nutrition* 113 (5): 1275–81. https://doi.org/10.1093/ajcn/nqaa381.

Ma, Yue, Yuan Zhang, Shu Li, Hongxi Yang, Huiping Li, Zhi Cao, Fusheng Xu, Li Sun, and Yaogang Wang. 2021. 'Sex Differences in Association Between Anti-Hypertensive Medications and Risk of COVID-19 in Middle-Aged and Older Adults'. *Drugs & Aging* 38 (10): 921–30. https://doi.org/10.1007/s40266-021-00886-y.

Mackay, A, D F Mackay, C A Celis-Morales, D M Lyall, S R Gray, N Sattar, J M R Gill, J P Pell, and J J Anderson. 2019. 'The Association between Driving Time and Unhealthy Lifestyles: A Cross-Sectional, General Population Study of 386 493 UK Biobank Participants'. *Journal of Public Health (Oxford, England)* 41 (3): 527–34. https://doi.org/10.1093/pubmed/fdy155.

Maiden, Martin C. J., Jane A. Bygraves, Edward Feil, Giovanna Morelli, Joanne E. Russell, Rachel Urwin, Qing Zhang, et al. 1998a. 'Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms'. *Proceedings of the National Academy of Sciences of the United States of America* 95 (6): 3140–45.

———. 1998b. 'Multilocus Sequence Typing: A Portable Approach to the Identification of Clones within Populations of Pathogenic Microorganisms'. *Proceedings of the National Academy of Sciences of the United States of America* 95 (6): 3140–45. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC19708/.

Maidstone, Robert, Simon G Anderson, David W Ray, Martin K Rutter, Hannah J Durrington, and John F Blaikley. 2021. 'Shift Work Is Associated with Positive COVID-19 Status in Hospitalised Patients'. *Thorax* 76 (6): 601–6. https://doi.org/10.1136/thoraxjnl-2020-216651.

Makowski, Dominique, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. 'BayestestR: Describing Effects and Their Uncertainty, Existence and Significance within the Bayesian Framework'. *Journal of Open Source Software* 4 (40): 1541. https://doi.org/10.21105/joss.01541.

Malavazos, Alexis Elias, Francesco Secchi, Sara Basilico, Gloria Capitanio, Sara Boveri, Valentina Milani, Carola Dubini, et al. 2022. 'Abdominal Obesity Phenotype Is Associated with COVID-19 Chest X-Ray Severity Score Better than BMI-Based Obesity'. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity* 27 (1): 345–59. https://doi.org/10.1007/s40519-021-01173-w.

Mallick, Himel, and Nengjun Yi. 2013. 'Bayesian Methods for High Dimensional Linear Models'. *Journal of Biometrics & Biostatistics* 1 (June): 005. https://doi.org/10.4172/2155-6180.S1-005.

Marchant, Joanna, Brendan Wren, and Julian Ketley. 2002. 'Exploiting Genome Sequence: Predictions for Mechanisms of Campylobacter Chemotaxis'. *Trends in Microbiology* 10 (4): 155–59. https://doi.org/10.1016/S0966-842X(02)02323-5.

Martin, Christopher A., David R. Jenkins, Jatinder S. Minhas, Laura J. Gray, Julian Tang, Caroline Williams, Shirley Sze, et al. 2020. 'Socio-Demographic Heterogeneity in the Prevalence of COVID-19 during Lockdown Is Associated with Ethnicity and Household Size: Results from an Observational Cohort Study'. *EClinicalMedicine* 25 (August). https://doi.org/10.1016/j.eclinm.2020.100466.

Mason, Llew, Jonathan Baxter, Peter Bartlett, and Marcus Frean. 1999. 'Boosting Algorithms as Gradient Descent'. In *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press. https://proceedings.neurips.cc/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html.

Mathur, Rohini, Christopher T. Rentsch, Caroline E. Morton, William J. Hulme, Anna Schultze, Brian MacKenna, Rosalind M. Eggo, et al. 2021. 'Ethnic Differences in SARS-CoV-2 Infection and COVID-19-Related Hospitalisation, Intensive Care Unit Admission, and Death in 17 Million Adults in England: An Observational Cohort Study Using the OpenSAFELY Platform'. *The Lancet* 397 (10286): 1711–24. https://doi.org/10.1016/S0140-6736(21)00634-6.

Matsuzawa, Yuji, Iichiro Shimomura, Tadashi Nakamura, Yoshiaki Keno, Kasuaki Kotani, and Katsuto Tokunaga. 1995. 'Pathophysiology and Pathogenesis of Visceral Fat Obesity'. *Obesity Research* 3 (S2): 187s–94. https://doi.org/10.1002/j.1550-8528.1995.tb00462.x.

McCoy, John, Carlos G. Wambier, Sergio Vano-Galvan, Jerry Shapiro, Rodney Sinclair, Paulo Müller Ramos, Kenneth Washenik, Murilo Andrade, Sabina Herrera, and Andy Goren. 2020. 'Racial Variations in COVID-19 Deaths May Be Due to Androgen Receptor Genetic Variants Associated with Prostate Cancer and Androgenetic Alopecia. Are Anti-Androgens a Potential Treatment for COVID-19?' *Journal of Cosmetic Dermatology*, April, 10.1111/jocd.13455. https://doi.org/10.1111/jocd.13455.

McIntyre, Alexa B. R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, et al. 2017. 'Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers'. *Genome Biology* 18 (September): 182. https://doi.org/10.1186/s13059-017-1299-7.

McQueenie, Ross, Hamish M. E. Foster, Bhautesh D. Jani, Srinivasa Vittal Katikireddi, Naveed Sattar, Jill P. Pell, Frederick K. Ho, et al. 2020. 'Multimorbidity, Polypharmacy, and COVID-19 Infection within the UK Biobank Cohort'. *PLOS ONE* 15 (8): e0238091. https://doi.org/10.1371/journal.pone.0238091.

Menardo, Fabrizio, Chloé Loiseau, Daniela Brites, Mireia Coscolla, Sebastian M. Gygli, Liliana K. Rutaihwa, Andrej Trauner, Christian Beisel, Sonia Borrell, and Sebastien Gagneux. 2018. 'Treemmer: A Tool to Reduce Large Phylogenetic Datasets with Minimal Loss of Diversity'. *BMC Bioinformatics* 19 (May): 164. https://doi.org/10.1186/s12859-018-2164-8.

Méric, Guillaume, Alan McNally, Alberto Pessia, Evangelos Mourkas, Ben Pascoe, Leonardos Mageiros, Minna Vehkala, Jukka Corander, and Samuel K Sheppard. 2018. 'Convergent Amino Acid Signatures in Polyphyletic Campylobacter Jejuni Subpopulations Suggest Human Niche Tropism'. *Genome Biology and Evolution* 10 (3): 763–74. https://doi.org/10.1093/gbe/evy026.

Molnar, Christoph. 2020. *Interpretable Machine Learning*. Lulu.com.

Mooney, Stephen J, Daniel J Westreich, and Abdulrahman M El-Sayed. 2015. 'Epidemiology in the Era of Big Data'. *Epidemiology (Cambridge, Mass.)* 26 (3): 390–94. https://doi.org/10.1097/EDE.0000000000000274.

Moore, Gordon E. 1965. 'Cramming More Components onto Integrated Circuits' 38 (8): 4.

Mouliou, Dimitra S., Ourania S. Kotsiou, and Konstantinos I. Gourgoulianis. 2021. 'Estimates of COVID-19 Risk Factors among Social Strata and Predictors for a Vulnerability to the Infection'. *International Journal of Environmental Research and Public Health* 18 (16): 8701. https://doi.org/10.3390/ijerph18168701.

Mourkas, Evangelos, Aidan J. Taylor, Guillaume Méric, Sion C. Bayliss, Ben Pascoe, Leonardos Mageiros, Jessica K. Calland, et al. 2020. 'Agricultural Intensification and the Evolution of Host Specialism in the Enteric Pathogen Campylobacter Jejuni'. *Proceedings*

of the National Academy of Sciences 117 (20): 11018–28. https://doi.org/10.1073/pnas.1917168117.

Mu, Yi, Isaac See, and Jonathan R. Edwards. 2019. 'Bayesian Model Averaging: Improved Variable Selection for Matched Case-Control Studies'. *Epidemiology, Biostatistics and Public Health* 16 (2): e13048. https://doi.org/10.2427/13048.

Mullner, Petra, Simon E. F. Spencer, Daniel J. Wilson, Geoff Jones, Alasdair D. Noble, Anne C. Midwinter, Julie M. Collins-Emerson, Philip Carter, Steve Hathaway, and Nigel P. French. 2009. 'Assigning the Source of Human Campylobacteriosis in New Zealand: A Comparative Genetic and Epidemiological Approach'. *Infection, Genetics and Evolution*, Includes papers from the Special Issue 'Parasitology in Mexico', 9 (6): 1311–19. https://doi.org/10.1016/j.meegid.2009.09.003.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. 'Definitions, Methods, and Applications in Interpretable Machine Learning'. *Proceedings of the National Academy of Sciences* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

Nachamkin, Irving, Ban Mishu Allos, and Tony Ho. 1998. 'Campylobacter Species and Guillain-Barré Syndrome'. *Clinical Microbiology Reviews* 11 (3): 555–67.

Naikare, Hemant, Kiran Palyada, Roger Panciera, Denver Marlow, and Alain Stintzi. 2006. 'Major Role for FeoB in Campylobacter Jejuni Ferrous Iron Acquisition, Gut Colonization, and Intracellular Survival'. *Infection and Immunity*, October. https://doi.org/10.1128/IAI.00052-06.

Neyman, J. 1937. 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability'. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236 (767): 333–80. https://www.jstor.org/stable/91337.

Nichols, Gordon L., Judith F. Richardson, Samuel K. Sheppard, Chris Lane, and Christophe Sarran. 2012. 'Campylobacter Epidemiology: A Descriptive Study Reviewing 1 Million Cases in England and Wales between 1989 and 2011'. *BMJ Open* 2 (4): e001179. https://doi.org/10.1136/bmjopen-2012-001179.

Nielsen, L.N., S.K. Sheppard, N.D. McCarthy, M.C.J. Maiden, H. Ingmer, and K.A. Krogfelt. 2010. 'MLST Clustering of Campylobacter Jejuni Isolates from Patients with Gastroenteritis, Reactive Arthritis and Guillain–Barré Syndrome'. *Journal of Applied Microbiology* 108 (2): 591–99. https://doi.org/10.1111/j.1365-2672.2009.04444.x.

Nikolenko, Sergey I., Anton I. Korobeynikov, and Max A. Alekseyev. 2013. 'BayesHammer: Bayesian Clustering for Error Correction in Single-Cell Sequencing'. *BMC Genomics* 14 (1): S7. https://doi.org/10.1186/1471-2164-14-S1-S7.

Nohra, Antoine, Alex Grinberg, Jonathan C. Marshall, Anne C. Midwinter, Julie M. Collins-Emerson, and Nigel P. French. 2020. 'Shifts in the Molecular Epidemiology of

Campylobacter Jejuni Infections in a Sentinel Region of New Zealand Following Implementation of Food Safety Interventions by the Poultry Industry'. *Applied and Environmental Microbiology* 86 (5). https://doi.org/10.1128/AEM.01753-19.

O'Brien, Sean M., and David B. Dunson. 2004. 'Bayesian Multivariate Logistic Regression'. *Biometrics* 60 (3): 739–46. https://doi.org/10.1111/j.0006-341X.2004.00224.x.

Ogata, Hiroaki, Masahiro Mori, Yujiro Jingushi, Hiroshi Matsuzaki, Katsuyuki Katahira, Akiko Ishimatsu, Aimi Enokizu-Ogawa, Kazuhito Taguchi, Atsushi Moriwaki, and Makoto Yoshida. 2021. 'Impact of Visceral Fat on the Prognosis of Coronavirus Disease 2019: An Observational Cohort Study'. *BMC Infectious Diseases* 21 (1): 1240. https://doi.org/10.1186/s12879-021-06958-z.

Ogden, Iain D., John F. Dallas, Marion MacRae, Ovidiu Rotariu, Kenny W. Reay, Malcolm Leitch, Ann P. Thomson, et al. 2009. 'Campylobacter Excreted into the Environment by Animal Sources: Prevalence, Concentration Shed, and Host Association'. *Foodborne Pathogens and Disease* 6 (10): 1161–70. https://doi.org/10.1089/fpd.2009.0327.

Oliveira, Manuela, Filipa Rocha Dias, and Constança Pomba. 2014. 'Biofilm and Fluoroquinolone Resistance of Canine Escherichia Coli Uropathogenic Isolates'. *BMC Research Notes* 7 (August): 499. https://doi.org/10.1186/1756-0500-7-499.

Olson, Donald R., Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. 2013. 'Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales'. *PLOS Computational Biology* 9 (10): e1003256. https://doi.org/10.1371/journal.pcbi.1003256.

Ow, Ghim Siong, Zhiqun Tang, and Vladimir A. Kuznetsov. 2016. 'Big Data and Computational Biology Strategy for Personalized Prognosis'. *Oncotarget* 7 (26): 40200–220. https://doi.org/10.18632/oncotarget.9571.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. 'APE: Analyses of Phylogenetics and Evolution in R Language'. *Bioinformatics* 20 (2): 289–90. https://doi.org/10.1093/bioinformatics/btg412.

Pascoe, Ben, Guillaume Méric, Susan Murray, Koji Yahara, Leonardos Mageiros, Ryan Bowen, Nathan H. Jones, et al. 2015. 'Enhanced Biofilm Formation and Multi-Host Transmission Evolve from Divergent Genetic Backgrounds in Campylobacter Jejuni'. *Environmental Microbiology* 17 (11): 4779–89. https://doi.org/10.1111/1462-2920.13051.

Patel, Aniruddh P., Manish D. Paranjpe, Nina P. Kathiresan, Manuel A. Rivas, and Amit V. Khera. 2020. 'Race, Socioeconomic Deprivation, and Hospitalization for COVID-19 in English Participants of a National Biobank'. *International Journal for Equity in Health* 19 (1): 114. https://doi.org/10.1186/s12939-020-01227-y.

Pearl, Judea. 2009. 'Causal Inference in Statistics: An Overview'. *Statistics Surveys* 3 (none): 96–146. https://doi.org/10.1214/09-SS057.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. 'Scikit-Learn: Machine Learning in Python'. *Journal of Machine Learning Research* 12: 2825–30.

Pei, Z., C. Burucoa, B. Grignon, S. Baqar, X. Z. Huang, D. J. Kopecko, A. L. Bourgeois, J. L. Fauchere, and M. J. Blaser. 1998. 'Mutation in the Peb1A Locus of Campylobacter Jejuni Reduces Interactions with Epithelial Cells and Intestinal Colonization of Mice'. *Infection and Immunity* 66 (3): 938–43. https://doi.org/10.1128/IAI.66.3.938-943.1998.

Pesaran, M. Hashem. 1990. 'Non-Nested Hypotheses'. In *Econometrics*, edited by John Eatwell, Murray Milgate, and Peter Newman, 167–73. The New Palgrave. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20570-7_24.

Peters, Sanne A. E., Stephen MacMahon, and Mark Woodward. 2021. 'Obesity as a Risk Factor for COVID-19 Mortality in Women and Men in the UK Biobank: Comparisons with Influenza/Pneumonia and Coronary Heart Disease'. *Diabetes, Obesity & Metabolism* 23 (1): 258–62. https://doi.org/10.1111/dom.14199.

Petersen, Antonia, Keno Bressem, Jakob Albrecht, Hans-Martin Thieß, Janis Vahldiek, Bernd Hamm, Marcus R. Makowski, Alexandra Niehues, Stefan M. Niehues, and Lisa C. Adams. 2020. 'The Role of Visceral Adiposity in the Severity of COVID-19: Highlights from a Unicenter Cross-Sectional Pilot Study in Germany'. *Metabolism* 110 (September): 154317. https://doi.org/10.1016/j.metabol.2020.154317.

Piddock, L. J. V. 2003. 'Fluoroquinolone Resistance in Campylobacter Species from Man and Animals: Detection of Mutations in Topoisomerase Genes'. *Journal of Antimicrobial Chemotherapy* 51 (1): 19–26. https://doi.org/10.1093/jac/dkg033.

Pohlman, John, and Dennis Leitner. 2003. 'A Comparison of Ordinary Least Squares and Logistic Regression'. *The Ohio Journal of Science* 103 (December).

Popkova, Elena G., and Kantoro Gulzat. 2020. 'Technological Revolution in the 21st Century: Digital Society vs. Artificial Intelligence'. In *The 21st Century from the Positions of Modern Science: Intellectual, Digital and Innovative Aspects*, edited by Elena G. Popkova and Bruno S. Sergi, 339–45. Lecture Notes in Networks and Systems. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32015-7_38.

Prats-Uribe, Albert, Junqing Xie, Daniel Prieto-Alhambra, and Irene Petersen. 2021. 'Smoking and COVID-19 Infection and Related Mortality: A Prospective Cohort Analysis of UK Biobank Data'. *Clinical Epidemiology* 13: 357–65. https://doi.org/10.2147/CLEP.S300597.

Preston, Jeremy, Ashley VanZeeland, and Daniel Pfeiffer. 2021. 'Innovation at Illumina: The Road to the $600 Human Genome'. https://www.nature.com/articles/d42473-021-00030-9.

Price, Alkes L., Noah A. Zaitlen, David Reich, and Nick Patterson. 2010. 'New Approaches to Population Stratification in Genome-Wide Association Studies'. *Nature Reviews Genetics* 11 (7): 459–63. https://doi.org/10.1038/nrg2813.

Price, Lance B., Leila G. Lackey, Rocio Vailes, and Ellen Silbergeld. 2007. 'The Persistence of Fluoroquinolone-Resistant Campylobacter in Poultry Production'. *Environmental Health Perspectives* 115 (7): 1035–39. https://doi.org/10.1289/ehp.10050.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. 'Inference of Population Structure Using Multilocus Genotype Data'. *Genetics* 155 (2): 945–59.

Prosperi, Mattia, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. 'Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare'. *Nature Machine Intelligence* 2 (7): 369–75. https://doi.org/10.1038/s42256-020-0197-y.

Puth, Marie-Therese, Markus Neuhäuser, and Graeme D. Ruxton. 2015. 'On the Variety of Methods for Calculating Confidence Intervals by Bootstrapping'. *Journal of Animal Ecology* 84 (4): 892–97. https://doi.org/10.1111/1365-2656.12382.

Quinlan, J. R. 1986. 'Induction of Decision Trees'. *Machine Learning* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Ragab, Dina, Haitham Salah Eldin, Mohamed Taeimah, Rasha Khattab, and Ramy Salem. 2020. 'The COVID-19 Cytokine Storm; What We Know So Far'. *Frontiers in Immunology* 11 (June): 1446. https://doi.org/10.3389/fimmu.2020.01446.

Raisi-Estabragh, Zahra, Celeste McCracken, Mae S Bethell, Jackie Cooper, Cyrus Cooper, Mark J Caulfield, Patricia B Munroe, Nicholas C Harvey, and Steffen E Petersen. 2020. 'Greater Risk of Severe COVID-19 in Black, Asian and Minority Ethnic Populations Is Not Explained by Cardiometabolic, Socioeconomic or Behavioural Factors, or by 25(OH)-Vitamin D Status: Study of 1326 Cases from the UK Biobank'. *Journal of Public Health* 42 (3): 451–60. https://doi.org/10.1093/pubmed/fdaa095.

Rashedi, Jalil, Behroz Mahdavi Poor, Vahid Asgharzadeh, Mahya Pourostadi, Hossein Samadi Kafil, Ali Vegari, Hamid Tayebi-Khosroshahi, and Mohammad Asgharzadeh. 2020. 'Risk Factors for COVID-19'. *Le Infezioni in Medicina* 28 (4): 469–74.

Raveendran, A. V., Rajeev Jayadevan, and S. Sashidharan. 2021. 'Long COVID: An Overview'. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 15 (3): 869–75. https://doi.org/10.1016/j.dsx.2021.04.007.

Razieh, Cameron, Francesco Zaccardi, Melanie J. Davies, Kamlesh Khunti, and Thomas Yates. 2020. 'Body Mass Index and the Risk of COVID-19 across Ethnic Groups: Analysis of UK Biobank'. *Diabetes, Obesity & Metabolism* 22 (10): 1953–54. https://doi.org/10.1111/dom.14125.

Reddy, N. R., S. K. Sathe, and D. K. Salunkhe. 1982. 'Phytates in Legumes and Cereals'. In *Advances in Food Research*, edited by C. O. Chichester, E. M. Mrak, and G. F. Stewart, 28:1–92. Academic Press. https://doi.org/10.1016/S0065-2628(08)60110-X.

Reece, Richard J., and Anthony Maxwell. 1991. 'DNA Gyrase: Structure and Function'. *Critical Reviews in Biochemistry and Molecular Biology* 26 (3–4): 335–75. https://doi.org/10.3109/10409239109114072.

Rhys, Hefin I. 2020. *Machine Learning with R, the Tidyverse, and Mlr*. Simon and Schuster.

Rizk, Guillaume, Dominique Lavenier, and Rayan Chikhi. 2013a. 'DSK: K-Mer Counting with Very Low Memory Usage'. *Bioinformatics* 29 (5): 652–53. https://doi.org/10.1093/bioinformatics/btt020.

———. 2013b. 'DSK: K-Mer Counting with Very Low Memory Usage'. *Bioinformatics (Oxford, England)* 29 (5): 652–53. https://doi.org/10.1093/bioinformatics/btt020.

Robert, Christian P. 2007. 'Model Choice'. In *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, edited by Christian P. Robert, 343–89. Springer Texts in Statistics. New York, NY: Springer. https://doi.org/10.1007/0-387-71599-1_7.

Roux, Francois, Emma Sproston, Ovidiu Rotariu, Marion MacRae, Samuel K. Sheppard, Paul Bessell, Alison Smith-Palmer, et al. 2013. 'Elucidating the Aetiology of Human Campylobacter Coli Infections'. *PLoS ONE* 8 (5). https://doi.org/10.1371/journal.pone.0064504.

Saarela, Mirka, and Susanne Jauhiainen. 2021. 'Comparison of Feature Importance Measures as Explanations for Classification Models'. *SN Applied Sciences* 3 (2): 272. https://doi.org/10.1007/s42452-021-04148-9.

Safizadeh, Fatemeh, Thi Ngoc Mai Nguyen, Hermann Brenner, and Ben Schöttker. 2021. 'Association of Renin-Angiotensin-Aldosterone System Inhibition with Covid-19 Hospitalization and All-Cause Mortality in the UK Biobank'. *British Journal of Clinical Pharmacology*, December. https://doi.org/10.1111/bcp.15192.

Sattar, Naveed, Frederick K. Ho, Jason MR. Gill, Nazim Ghouri, Stuart R. Gray, Carlos A. Celis-Morales, S. Vittal Katikireddi, et al. 2020. 'BMI and Future Risk for COVID-19 Infection and Death across Sex, Age and Ethnicity: Preliminary Findings from UK Biobank'. *Diabetes & Metabolic Syndrome* 14 (5): 1149–51. https://doi.org/10.1016/j.dsx.2020.06.060.

Sauerbrei, Willi, Aris Perperoglou, Matthias Schmid, Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, et al. 2020. 'State of the Art in Selection of Variables and Functional Forms in Multivariable Analysis—Outstanding Issues'. *Diagnostic and Prognostic Research* 4 (1): 3. https://doi.org/10.1186/s41512-020-00074-3.

Scanlan, Eoin, Laura Ardill, Matthew V. X. Whelan, Claire Shortt, Jarlath E. Nally, Billy Bourke, and Tadhg Ó Cróinín. 2017. 'Relaxation of DNA Supercoiling Leads to Increased Invasion of Epithelial Cells and Protein Secretion by Campylobacter Jejuni'. *Molecular Microbiology* 104 (1): 92–104. https://doi.org/10.1111/mmi.13614.

Schrider, Daniel R., and Andrew D. Kern. 2018. 'Supervised Machine Learning for Population Genetics: A New Paradigm'. *Trends in Genetics* 34 (4): 301–12. https://doi.org/10.1016/j.tig.2017.12.005.

Sears, Ann, Michael G. Baker, Nick Wilson, Jonathan Marshall, Petra Muellner, Donald M. Campbell, Robin J. Lake, and Nigel P. French. 2011. 'Marked Campylobacteriosis Decline after Interventions Aimed at Poultry, New Zealand'. *Emerging Infectious Diseases* 17 (6): 1007–15. https://doi.org/10.3201/eid1706.101272.

Sejnowski, Terrence J. 2018. *The Deep Learning Revolution*. MIT Press.

Service, Robert F. 2006. 'The Race for the $1000 Genome'. *Science* 311 (5767): 1544–46. https://doi.org/10.1126/science.311.5767.1544.

Shalf, John. 2020. 'The Future of Computing beyond Moore's Law'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378 (2166): 20190061. https://doi.org/10.1098/rsta.2019.0061.

Sheehan, Sara, and Yun S. Song. 2016. 'Deep Learning for Population Genetic Inference'. *PLOS Computational Biology* 12 (3): e1004845. https://doi.org/10.1371/journal.pcbi.1004845.

Shen, Mingwang, Zhihang Peng, Yuming Guo, Libin Rong, Yan Li, Yanni Xiao, Guihua Zhuang, and Lei Zhang. 2020. 'Assessing the Effects of Metropolitan-Wide Quarantine on the Spread of COVID-19 in Public Space and Households'. *International Journal of Infectious Diseases* 96 (July): 503–5. https://doi.org/10.1016/j.ijid.2020.05.019.

Sheppard, Samuel K., Lu Cheng, Guillaume Méric, Caroline P. A. de Haan, Ann-Katrin Llarena, Pekka Marttinen, Ana Vidal, et al. 2014. 'Cryptic Ecology among Host Generalist Campylobacter Jejuni in Domestic Animals'. *Molecular Ecology* 23 (10): 2442–51. https://doi.org/10.1111/mec.12742.

Sheppard, Samuel K., Lu Cheng, Guillaume Méric, Caroline P. A. de Haan, Ann-Katrin Llarena, Pekka Marttinen, Ana Vidal, et al. 2014. 'Cryptic Ecology among Host Generalist Campylobacter Jejuni in Domestic Animals'. *Molecular Ecology* 23 (10): 2442–51. https://doi.org/10.1111/mec.12742.

Sheppard, Samuel K., Frances M. Colles, Noel D. McCARTHY, Norval J. C. Strachan, Iain D. Ogden, Ken J. Forbes, John F. Dallas, and Martin C. J. Maiden. 2011a. 'Niche Segregation and Genetic Structure of Campylobacter Jejuni Populations from Wild and Agricultural Host Species'. *Molecular Ecology* 20 (16): 3484–90. https://doi.org/10.1111/j.1365-294X.2011.05179.x.

———. 2011b. ‘Niche Segregation and Genetic Structure of Campylobacter Jejuni Populations from Wild and Agricultural Host Species’. *Molecular Ecology* 20 (16): 3484–90. https://doi.org/10.1111/j.1365-294X.2011.05179.x.

Sheppard, Samuel K., Frances Colles, Judith Richardson, Alison J. Cody, Richard Elson, Andrew Lawson, Geraldine Brick, et al. 2010a. ‘Host Association of Campylobacter Genotypes Transcends Geographic Variation’. *Applied and Environmental Microbiology* 76 (15): 5269–77. https://doi.org/10.1128/AEM.00124-10.

———. 2010b. ‘Host Association of Campylobacter Genotypes Transcends Geographic Variation’. *Applied and Environmental Microbiology*, August. https://doi.org/10.1128/AEM.00124-10.

Sheppard, Samuel K., John F. Dallas, Marion MacRae, Noel D. McCarthy, E. L. Sproston, F. J. Gormley, Norval J. C. Strachan, Iain D. Ogden, Martin C. J. Maiden, and Ken J. Forbes. 2009a. ‘Campylobacter Genotypes from Food Animals, Environmental Sources and Clinical Disease in Scotland 2005/6’. *International Journal of Food Microbiology* 134 (1–2): 96–103. https://doi.org/10.1016/j.ijfoodmicro.2009.02.010.

———. 2009b. ‘Campylobacter Genotypes from Food Animals, Environmental Sources and Clinical Disease in Scotland 2005/6’. *International Journal of Food Microbiology*, Food Micro 2008 “Evolving Microbial Food Safety and Quality” 1–4 September 2008, Aberdeen, Scotland, UK, 134 (1): 96–103. https://doi.org/10.1016/j.ijfoodmicro.2009.02.010.

Sheppard, Samuel K., John F. Dallas, Norval J. C. Strachan, Marian MacRae, Noel D. McCarthy, Daniel J. Wilson, Fraser J. Gormley, et al. 2009a. ‘Campylobacter Genotyping to Determine the Source of Human Infection’. *Clinical Infectious Diseases* 48 (8): 1072–78. https://doi.org/10.1086/597402.

———. 2009b. ‘Campylobacter Genotyping to Determine the Source of Human Infection’. *Clinical Infectious Diseases* 48 (8): 1072–78. https://doi.org/10.1086/597402.

Sheppard, Samuel K., John F. Dallas, Daniel J. Wilson, Norval J. C. Strachan, Noel D. McCarthy, Keith A. Jolley, Frances M. Colles, et al. 2010. ‘Evolution of an Agriculture-Associated Disease Causing Campylobacter Coli Clade: Evidence from National Surveillance Data in Scotland’. *PLOS ONE* 5 (12): e15708. https://doi.org/10.1371/journal.pone.0015708.

Sheppard, Samuel K., Xavier Didelot, Guillaume Meric, Alicia Torralbo, Keith A. Jolley, David J. Kelly, Stephen D. Bentley, Martin C. J. Maiden, Julian Parkhill, and Daniel Falush. 2013a. ‘Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter’. *Proceedings of the National Academy of Sciences* 110 (29): 11923–27. https://doi.org/10.1073/pnas.1305559110.

———. 2013b. ‘Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter’. *Proceedings of the National Academy of Sciences* 110 (29): 11923–27. https://doi.org/10.1073/pnas.1305559110.

———. 2013c. ‘Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter’. *Proceedings of the National Academy of Sciences* 110 (29): 11923–27. https://doi.org/10.1073/pnas.1305559110.

Sheppard, Samuel K., Keith A. Jolley, and Martin C. J. Maiden. 2012. ‘A Gene-By-Gene Approach to Bacterial Population Genomics: Whole Genome MLST of Campylobacter’. *Genes* 3 (2): 261–77. https://doi.org/10.3390/genes3020261.

Sheppard, Samuel K., and Martin C.J. Maiden. 2015. ‘The Evolution of Campylobacter Jejuni and Campylobacter Coli’. *Cold Spring Harbor Perspectives in Biology* 7 (8). https://doi.org/10.1101/cshperspect.a018119.

Shi, Ning-Zhong, and Jian Tao. 2008. *Statistical Hypothesis Testing: Theory and Methods*. World Scientific.

Shortt, Claire, Eoin Scanlan, Amber Hilliard, Chiara E. Cotroneo, Billy Bourke, and Tadhg Ó Cróinín. 2016. ‘DNA Supercoiling Regulates the Motility of Campylobacter Jejuni and Is Altered by Growth in the Presence of Chicken Mucus’. *MBio*, September. https://doi.org/10.1128/mBio.01227-16.

Simonsen, Lone, Julia R. Gog, Don Olson, and Cécile Viboud. 2016. ‘Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems’. *The Journal of Infectious Diseases* 214 (suppl_4): S380–85. https://doi.org/10.1093/infdis/jiw376.

Skelly, Andrea C., Joseph R. Dettori, and Erika D. Brodt. 2012. ‘Assessing Bias: The Importance of Considering Confounding’. *Evidence-Based Spine-Care Journal* 3 (1): 9–12. https://doi.org/10.1055/s-0031-1298595.

Sokolowska, Milena, Zuzanna M. Lukasik, Ioana Agache, Cezmi A. Akdis, Deniz Akdis, Mübeccel Akdis, Weronika Barcik, et al. 2020. ‘Immunology of COVID-19: Mechanisms, Clinical Outcome, Diagnostics, and Perspectives—A Report of the European Academy of Allergy and Clinical Immunology (EAACI)’. *Allergy* 75 (10): 2445–76. https://doi.org/10.1111/all.14462.

Sönmez, Sevil, Jessica Wiitala, and Yorgos Apostolopoulos. 2019. ‘How Complex Travel, Tourism, and Transportation Networks Influence Infectious Disease Movement in a Borderless World’. *Handbook of Globalisation and Tourism*, December. https://www.elgaronline.com/view/edcoll/9781786431288/9781786431288.00015.xml.

Sproston, Emma L., Helen M. L. Wimalarathna, and Samuel K. Sheppard. 2018. ‘Trends in Fluoroquinolone Resistance in Campylobacter’. *Microbial Genomics* 4 (8): e000198. https://doi.org/10.1099/mgen.0.000198.

Statnikov, Alexander, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin J. Blaser, Constantin F. Aliferis, and Alexander V. Alekseyenko. 2013. 'A Comprehensive Evaluation of Multicategory Classification Methods for Microbiomic Data'. *Microbiome* 1 (1): 11. https://doi.org/10.1186/2049-2618-1-11.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. 'Big Data: Astronomical or Genomical?' *PLOS Biology* 13 (7): e1002195. https://doi.org/10.1371/journal.pbio.1002195.

Stoltzfus, Jill C. 2011. 'Logistic Regression: A Brief Primer'. *Academic Emergency Medicine* 18 (10): 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x.

Stone, M. 1974. 'Cross-Validatory Choice and Assessment of Statistical Predictions'. *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (2): 111–47. https://www.jstor.org/stable/2984809.

Strachan, Norval J. C., Fraser J. Gormley, Ovidiu Rotariu, Iain D. Ogden, Gordon Miller, Geoff M. Dunn, Samuel K. Sheppard, et al. 2009. 'Attribution of Campylobacter Infections in Northeast Scotland to Specific Sources by Use of Multilocus Sequence Typing'. *J Infect Dis* 199 (8): 1205–8. https://doi.org/10.1086/597417.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age'. *PLOS Medicine* 12 (3): e1001779. https://doi.org/10.1371/journal.pmed.1001779.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning, Second Edition: An Introduction*. MIT Press.

Suyanto, S. A. N., B. Siregar, E. B. Nababan, and H. A. Fikri. 2020. 'Classification of Infection Type Based on Leukocytes Examination Results Using K-Nearest Neighbor'. *Journal of Physics: Conference Series* 1566 (1): 012130. https://doi.org/10.1088/1742-6596/1566/1/012130.

Szabo, Peter A., Pranay Dogra, Joshua I. Gray, Steven B. Wells, Thomas J. Connors, Stuart P. Weisberg, Izabela Krupska, et al. 2021. 'Longitudinal Profiling of Respiratory and Systemic Immune Responses Reveals Myeloid Cell-Driven Lung Inflammation in Severe COVID-19'. *Immunity* 54 (4): 797-814.e6. https://doi.org/10.1016/j.immuni.2021.03.005.

Sze, Shirley, Daniel Pan, Clareece R. Nevill, Laura J. Gray, Christopher A. Martin, Joshua Nazareth, Jatinder S. Minhas, et al. 2020. 'Ethnicity and Clinical Outcomes in COVID-19: A Systematic Review and Meta-Analysis'. *EClinicalMedicine* 29 (December): 100630. https://doi.org/10.1016/j.eclinm.2020.100630.

Tanaka, Kari J., Saemee Song, Kevin Mason, and Heather W. Pinkett. 2018. 'Selective Substrate Uptake: The Role of ATP-Binding Cassette (ABC) Importers in Pathogenesis'.

Biochimica Et Biophysica Acta. Biomembranes 1860 (4): 868–77. https://doi.org/10.1016/j.bbamem.2017.08.011.

Tang, Yizhi, Orhan Sahin, Nada Pavlovic, Jeff LeJeune, James Carlson, Zuowei Wu, Lei Dai, and Qijing Zhang. 2017. 'Rising Fluoroquinolone Resistance in Campylobacter Isolated from Feedlot Cattle in the United States'. *Scientific Reports* 7 (1): 494. https://doi.org/10.1038/s41598-017-00584-z.

Taveirne, Michael E., Casey M. Theriot, Jonathan Livny, and Victor J. DiRita. 2013. 'The Complete Campylobacter Jejuni Transcriptome during Colonization of a Natural Host Determined by RNAseq'. *PLOS ONE* 8 (8): e73586. https://doi.org/10.1371/journal.pone.0073586.

The CRyPTIC Consortium. 2021. 'Genome-Wide Association Studies of Global Mycobacterium Tuberculosis Resistance to Thirteen Antimicrobials in 10,228 Genomes'. bioRxiv. https://doi.org/10.1101/2021.09.14.460272.

'The European Union One Health 2018 Zoonoses Report'. 2019. *EFSA Journal* 17 (12): e05926. https://doi.org/10.2903/j.efsa.2019.5926.

The Severe Covid-19 GWAS Group. 2020. 'Genomewide Association Study of Severe Covid-19 with Respiratory Failure'. *New England Journal of Medicine* 383 (16): 1522–34. https://doi.org/10.1056/NEJMoa2020283.

Thépault, Amandine, Guillaume Méric, Katell Rivoal, Ben Pascoe, Leonardos Mageiros, Fabrice Touzain, Valérie Rose, Véronique Béven, Marianne Chemaly, and Samuel K. Sheppard. 2017. 'Genome-Wide Identification of Host-Segregating Epidemiological Markers for Source Attribution in Campylobacter Jejuni'. *Applied and Environmental Microbiology* 83 (7). https://doi.org/10.1128/AEM.03085-16.

Tom, Jennifer A., Jens Reeder, William F. Forrest, Robert R. Graham, Julie Hunkapiller, Timothy W. Behrens, and Tushar R. Bhangale. 2017. 'Identifying and Mitigating Batch Effects in Whole Genome Sequencing Data'. *BMC Bioinformatics* 18 (1): 351. https://doi.org/10.1186/s12859-017-1756-z.

Torre Díez, Isabel de la, Héctor Merino Cosgaya, Begoña Garcia-Zapirain, and Miguel López-Coronado. 2016. 'Big Data in Health: A Literature Review from the Year 2005'. *Journal of Medical Systems* 40 (9): 209. https://doi.org/10.1007/s10916-016-0565-7.

Torres, Rita Tinoco, João Carvalho, Joana Fernandes, Josman D. Palmeira, Mónica V. Cunha, and Carlos Fonseca. 2021. 'Mapping the Scientific Knowledge of Antimicrobial Resistance in Food-Producing Animals'. *One Health* 13 (December): 100324. https://doi.org/10.1016/j.onehlt.2021.100324.

Tran Kiem, Cécile, Paolo Bosetti, Juliette Paireau, Pascal Crépey, Henrik Salje, Noémie Lefrancq, Arnaud Fontanet, et al. 2021. 'SARS-CoV-2 Transmission across Age Groups in France and Implications for Control'. *Nature Communications* 12 (1): 6895. https://doi.org/10.1038/s41467-021-27163-1.

Travaglio, Marco, Yizhou Yu, Rebeka Popovic, Liza Selley, Nuno Santos Leal, and Luis Miguel Martins. 2021. 'Links between Air Pollution and COVID-19 in England'. *Environmental Pollution* 268 (January): 115859. https://doi.org/10.1016/j.envpol.2020.115859.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. 'Genome-Wide Association Studies'. *Nature Reviews Methods Primers* 1 (1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Van Tienderen, Peter H. 1991. 'Evolution of Generalists and Specialist in Spatially Heterogeneous Environments'. *Evolution* 45 (6): 1317–31. https://doi.org/10.2307/2409882.

Vatansever, Hafize Seda, and Eda Becer. 2020. 'Relationship between IL-6 and COVID-19: To Be Considered during Treatment'. *Future Virology* 15 (12): 817–22. https://doi.org/10.2217/fvl-2020-0168.

Velayudhan, Jyoti, and David J. Kelly. 2002. 'Analysis of Gluconeogenic and Anaplerotic Enzymes in Campylobacter Jejuni: An Essential Role for Phosphoenolpyruvate Carboxykinase'. *Microbiology (Reading, England)* 148 (Pt 3): 685–94. https://doi.org/10.1099/00221287-148-3-685.

Ventola, C. Lee. 2015. 'The Antibiotic Resistance Crisis'. *Pharmacy and Therapeutics* 40 (4): 277–83. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/.

Vlaming, Ronald de, and Patrick J. F. Groenen. 2015. 'The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics'. *BioMed Research International* 2015 (July): e143712. https://doi.org/10.1155/2015/143712.

Vries, Stefan P. de, Srishti Gupta, Abiyad Baig, Elli Wright, Amy Wedley, Annette Nygaard Jensen, Lizeth LaCharme Lora, et al. 2017. 'Genome-Wide Fitness Analyses of the Foodborne Pathogen Campylobacter Jejuni in in Vitro and in Vivo Models'. *Scientific Reports* 7 (April): 1251. https://doi.org/10.1038/s41598-017-01133-4.

Wagenmakers, Eric-Jan, Michael Lee, Tom Lodewyckx, and Geoffrey J. Iverson. 2008. 'Bayesian Versus Frequentist Inference'. In *Bayesian Evaluation of Informative Hypotheses*, edited by Herbert Hoijtink, Irene Klugkist, and Paul A. Boelen, 181–207. Statistics for Social and Behavioral Sciences. New York, NY: Springer. https://doi.org/10.1007/978-0-387-09612-4_9.

Wang, Qi, Yue Ma, Kun Zhao, and Yingjie Tian. 2022. 'A Comprehensive Survey of Loss Functions in Machine Learning'. *Annals of Data Science* 9 (2): 187–212. https://doi.org/10.1007/s40745-020-00253-5.

Wang, Yongjun, Yang Yang, Lina Ren, Yuan Shao, Weiqun Tao, and Xi-jian Dai. 2021. 'Preexisting Mental Disorders Increase the Risk of COVID-19 Infection and Associated

Mortality'. *Frontiers in Public Health* 9. https://www.frontiersin.org/article/10.3389/fpubh.2021.684112.

Waskom, Michael L. 2021. 'Seaborn: Statistical Data Visualization'. *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.

Webb, Geoffrey I. 2010. 'Naïve Bayes'. In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 713–14. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_576.

Weis, Caroline, Aline Cuénod, Bastian Rieck, Olivier Dubuis, Susanne Graf, Claudia Lang, Michael Oberle, et al. 2022. 'Direct Antimicrobial Resistance Prediction from Clinical MALDI-TOF Mass Spectra Using Machine Learning'. *Nature Medicine* 28 (1): 164–74. https://doi.org/10.1038/s41591-021-01619-9.

Whelan, Matthew V. X., Laura Ardill, Kentaro Koide, Chie Nakajima, Yasuhiko Suzuki, Jeremy C. Simpson, and Tadhg Ó Cróinín. 2019. 'Acquisition of Fluoroquinolone Resistance Leads to Increased Biofilm Formation and Pathogenicity in Campylobacter Jejuni'. *Scientific Reports* 9 (1): 18216. https://doi.org/10.1038/s41598-019-54620-1.

Wiemken, Timothy L., and Robert R. Kelley. 2020. 'Machine Learning in Epidemiology and Health Outcomes Research'. *Annual Review of Public Health* 41 (April): 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437.

Willmott, Christopher J. R., Susan E. Critchlow, Ian C. Eperon, and Anthony Maxwell. 1994. 'The Complex of DNA Gyrase and Quinolone Drugs with DNA Forms a Barrier to Transcription by RNA Polymerase'. *Journal of Molecular Biology* 242 (4): 351–63. https://doi.org/10.1006/jmbi.1994.1586.

Wilson, Daniel J., Edith Gabriel, Andrew J. H. Leatherbarrow, John Cheesbrough, Steven Gee, Eric Bolton, Andrew Fox, Paul Fearnhead, C. Anthony Hart, and Peter J. Diggle. 2008a. 'Tracing the Source of Campylobacteriosis'. *PLOS Genetics* 4 (9): e1000203. https://doi.org/10.1371/journal.pgen.1000203.

———. 2008b. 'Tracing the Source of Campylobacteriosis'. *PLOS Genetics* 4 (9): e1000203. https://doi.org/10.1371/journal.pgen.1000203.

Wolpert, David H. 1996. 'The Lack of A Priori Distinctions Between Learning Algorithms'. *Neural Computation* 8 (7): 1341–90. https://doi.org/10.1162/neco.1996.8.7.1341.

Wong, Ho Ting, Qian Yin, Ying Qi Guo, Kristen Murray, Dong Hau Zhou, and Diana Slade. 2015. 'Big Data as a New Approach in Emergency Medicine Research'. *Journal of Acute Disease* 4 (3): 178–79. https://doi.org/10.1016/j.joad.2015.04.003.

Wong, Kenneth Chi-Yin, Yong Xiang, Liangying Yin, and Hon-Cheong So. 2021. 'Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach'. *JMIR Public Health and Surveillance* 7 (9): e29544. https://doi.org/10.2196/29544.

Wong, Zoie S. Y., Jiaqi Zhou, and Qingpeng Zhang. 2019. 'Artificial Intelligence for Infectious Disease Big Data Analytics'. *Infection, Disease & Health* 24 (1): 44–48. https://doi.org/10.1016/j.idh.2018.10.002.

Wood, Derrick E., and Steven L. Salzberg. 2014. 'Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments'. *Genome Biology* 15 (3): 1–12. https://doi.org/10.1186/gb-2014-15-3-r46.

Woodcock, Dan J., Peter Krusche, Norval J. C. Strachan, Ken J. Forbes, Frederick M. Cohan, Guillaume Méric, and Samuel K. Sheppard. 2017. 'Genomic Plasticity and Rapid Host Switching Can Promote the Evolution of Generalism: A Case Study in the Zoonotic Pathogen Campylobacter'. *Scientific Reports* 7 (1): 1–13. https://doi.org/10.1038/s41598-017-09483-9.

Woodward, Mark, Sanne A. E. Peters, and Katie Harris. 2021. 'Social Deprivation as a Risk Factor for COVID-19 Mortality among Women and Men in the UK Biobank: Nature of Risk and Context Suggests That Social Interventions Are Essential to Mitigate the Effects of Future Pandemics'. *J Epidemiol Community Health* 75 (11): 1050–55. https://doi.org/10.1136/jech-2020-215810.

Woolf, Barnet. 1957. 'The Log Likelihood Ratio Test (the G-Test)'. *Annals of Human Genetics* 21 (4): 397–409. https://doi.org/10.1111/j.1469-1809.1972.tb00293.x.

Woolhouse, M. E., L. H. Taylor, and D. T. Haydon. 2001. 'Population Biology of Multihost Pathogens'. *Science (New York, N.Y.)* 292 (5519): 1109–12. https://doi.org/10.1126/science.1059026.

World Health Organization. 1992. 'The ICD-10 Classification of Mental and Behavioural Disorders : Clinical Descriptions and Diagnostic Guidelines'. World Health Organization. https://apps.who.int/iris/handle/10665/37958.

Wright, Raymond E. 1995. 'Logistic Regression'. In *Reading and Understanding Multivariate Statistics*, 217–44. Washington, DC, US: American Psychological Association.

Xia, Jing, Jinji Pang, Yizhi Tang, Zuowei Wu, Lei Dai, Kritika Singh, Changyun Xu, et al. 2019. 'High Prevalence of Fluoroquinolone-Resistant Campylobacter Bacteria in Sheep and Increased Campylobacter Counts in the Bile and Gallbladders of Sheep Medicated with Tetracycline in Feed'. *Applied and Environmental Microbiology* 85 (11): e00008-19. https://doi.org/10.1128/AEM.00008-19.

Xiang, Yong, Kenneth Chi-Yin Wong, and Hon-Cheong So. 2021. 'Exploring Drugs and Vaccines Associated with Altered Risks and Severity of COVID-19: A UK Biobank Cohort Study of All ATC Level-4 Drug Categories Reveals Repositioning Opportunities'. *Pharmaceutics* 13 (9): 1514. https://doi.org/10.3390/pharmaceutics13091514.

Xu, Shuo. 2018. 'Bayesian Naïve Bayes Classifiers to Text Classification'. *Journal of Information Science* 44 (1): 48–59. https://doi.org/10.1177/0165551516677946.

Yahara, Koji, Guillaume Méric, Aidan J. Taylor, Stefan P. W. de Vries, Susan Murray, Ben Pascoe, Leonardos Mageiros, et al. 2017. 'Genome-Wide Association of Functional Traits Linked with Campylobacter Jejuni Survival from Farm to Fork'. *Environmental Microbiology* 19 (1): 361–80. https://doi.org/10.1111/1462-2920.13628.

Yahara, Koji, Guillaume Méric, Aidan J. Taylor, Stefan P. W. de Vries, Susan Murray, Ben Pascoe, Leonardos Mageiros, et al. 2017. 'Genome-Wide Association of Functional Traits Linked with Campylobacter Jejuni Survival from Farm to Fork'. *Environmental Microbiology* 19 (1): 361–80. https://doi.org/10.1111/1462-2920.13628.

Yang, Qiao, Jixi Li, Zhijia Zhang, Xiaocheng Wu, Tongquan Liao, Shiyong Yu, Zaichun You, et al. 2021. 'Clinical Characteristics and a Decision Tree Model to Predict Death Outcome in Severe COVID-19 Patients'. *BMC Infectious Diseases* 21 (1): 783. https://doi.org/10.1186/s12879-021-06478-w.

Yates, Thomas, Cameron Razieh, Francesco Zaccardi, Alex V. Rowlands, Samuel Seidu, Melanie J. Davies, and Kamlesh Khunti. 2021. 'Obesity, Walking Pace and Risk of Severe COVID-19 and Mortality: Analysis of UK Biobank'. *International Journal of Obesity* 45 (5): 1155–59. https://doi.org/10.1038/s41366-021-00771-z.

Yimer, Belay Birlie, Martin Otava, Teshome Degefa, Delenasaw Yewhalaw, and Ziv Shkedy. 2021. 'Bayesian Model Averaging in Longitudinal Studies Using Bayesian Variable Selection Methods'. *Communications in Statistics - Simulation and Computation* 0 (0): 1–18. https://doi.org/10.1080/03610918.2021.1914088.

Yong, Shee Chien, Pietro Roversi, James Lillington, Fernanda Rodriguez, Martin Krehenbrink, Oliver B. Zeldin, Elspeth F. Garman, Susan M. Lea, and Ben C. Berks. 2014. 'A Complex Iron-Calcium Cofactor Catalyzing Phosphotransfer Chemistry'. *Science*, September. https://doi.org/10.1126/science.1254237.

Yoshikawa, Masahiro, and Kensuke Asaba. 2021. 'Educational Attainment Decreases the Risk of COVID-19 Severity in the European Population: A Two-Sample Mendelian Randomization Study'. *Frontiers in Public Health* 9: 673451. https://doi.org/10.3389/fpubh.2021.673451.

Young, Bernadette C., Chieh-Hsi Wu, Jane Charlesworth, Sarah Earle, James R. Price, N. Claire Gordon, Kevin Cole, et al. 2021. 'Antimicrobial Resistance Determinants Are Associated with Staphylococcus Aureus Bacteraemia and Adaptation to the Healthcare Environment: A Bacterial Genome-Wide Association Study'. *Microbial Genomics* 7 (11): 000700. https://doi.org/10.1099/mgen.0.000700.

Zhang, Jin, Min Chen, Yangjun Wen, Yin Zhang, Yunan Lu, Shengmeng Wang, and Juncong Chen. 2021. 'A Fast Multi-Locus Ridge Regression Algorithm for High-Dimensional Genome-Wide Association Studies'. *Frontiers in Genetics* 12. https://www.frontiersin.org/article/10.3389/fgene.2021.649196.

Zhang, Shaokang, Shaoting Li, Weidong Gu, Henk den Bakker, Dave Boxrud, Angie Taylor, Chandler Roe, et al. 2019. 'Zoonotic Source Attribution of Salmonella Enterica Serotype Typhimurium Using Genomic Surveillance Data, United States'. *Emerging Infectious Diseases* 25 (1): 82–91. https://doi.org/10.3201/eid2501.180835.

Zhang, Shichao, Debo Cheng, Rongyao Hu, and Zhenyun Deng. 2018. 'Supervised Feature Selection Algorithm via Discriminative Ridge Regression'. *World Wide Web* 21 (6): 1545–62. https://doi.org/10.1007/s11280-017-0502-9.

Zhang, Yu-han, Xiao-fei Hu, Jie-chao Ma, Xian-qi Wang, Hao-ran Luo, Zi-feng Wu, Shu Zhang, et al. 2021. 'Clinical Applicable AI System Based on Deep Learning Algorithm for Differentiation of Pulmonary Infectious Disease'. *Frontiers in Medicine* 8. https://www.frontiersin.org/article/10.3389/fmed.2021.753055.

Zhou, Fei, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, et al. 2020. 'Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: A Retrospective Cohort Study'. *The Lancet* 395 (10229): 1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3.

Zhou, Xiang, and Matthew Stephens. 2012. 'Genome-Wide Efficient Mixed Model Analysis for Association Studies'. *Nature Genetics* 44 (7): 821–24. https://doi.org/10.1038/ng.2310.

Zhu, Zhaozhong, Kohei Hasegawa, Baoshan Ma, Michimasa Fujiogi, Carlos A. Camargo, and Liming Liang. 2020a. 'Association of Asthma and Its Genetic Predisposition with the Risk of Severe COVID-19'. *The Journal of Allergy and Clinical Immunology* 146 (2): 327-329.e4. https://doi.org/10.1016/j.jaci.2020.06.001.

———. 2020b. 'Association of Obesity and Its Genetic Predisposition with the Risk of Severe COVID-19: Analysis of Population-Based Cohort Data'. *Metabolism: Clinical and Experimental* 112 (November): 154345. https://doi.org/10.1016/j.metabol.2020.154345.

Zimin, Aleksey V., and Steven L. Salzberg. 2020. 'The Genome Polishing Tool POLCA Makes Fast and Accurate Corrections in Genome Assemblies'. *PLOS Computational Biology* 16 (6): e1007981. https://doi.org/10.1371/journal.pcbi.1007981.

Zou, Xiaonan, Yong Hu, Zhewen Tian, and Kaiyuan Shen. 2019. 'Logistic Regression Model Optimization and Case Analysis'. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 135–39. https://doi.org/10.1109/ICCSNT47585.2019.8962457.

Zulkower, Valentin, and Susan Rosser. 2020. 'DNA Features Viewer: A Sequence Annotation Formatting and Plotting Library for Python'. *Bioinformatics* 36 (15): 4350–52. https://doi.org/10.1093/bioinformatics/btaa213.